

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Production and Perception of non-native English in China

Focus on Sociophonetic Variation by Humans and Artificial Agents

D i s s e r t a t i o n

zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)

vorgelegt der Philosophischen Fakultät der
Technischen Universität Chemnitz

von Herrn Sven Albrecht

Erstgutachter: Prof. Dr. Josef Schmied
Zweitgutachter: Prof. Dr. Christina Sanchez-Stockhammer
Drittgutachter: Prof. Chenguang Chang, Ph.D.

Tag der Einreichung: 30.10.2023
Tag der Verteidigung: 12.12.2023
Veröffentlichungsjahr: 2024

Zitierlink: https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa2-888166

Contents

List of Figures **V**

List of Tables **VII**

List of Abbreviations **IX**

Acknowledgements **XI**

1 Introduction **1**
1.1 Embodied Digital Technologies and Artificial Agents . . . 2
1.2 Intentionality and Credibility . . . 2
1.2.1 Intentionality . . . 2
1.2.2 Credibility . . . 4
1.3 Sociolinguistic Theory . . . 5
1.3.1 Variationist Sociolinguistics . . . 5
1.3.1.1 The Research Framework . . . 6
1.3.1.2 Common Variables . . . 7
1.3.2 World Englishes . . . 8
1.3.2.1 Kachru's Concentric Three Circles . . . 10
1.3.2.2 Schneider's Dynamic Model . . . 10
1.3.2.3 On the Native Speaker . . . 12
1.4 The Linguistic Landscape of China . . . 12
1.5 Creating a Research Space . . . 14
1.6 Methodological Considerations . . . 15
1.6.1 Data Collection . . . 15
1.6.2 Data Collection and COVID-19 . . . 16
1.6.3 Data Transcription . . . 17
1.6.4 Automated Analysis . . . 20
1.6.5 Vowel Formant Normalization . . . 22
1.6.6 Vowel Length Normalization . . . 23
1.6.7 Text-to-Speech System . . . 24
1.7 Synopses . . . 25

2 Current research on the linguistic features of Chinese English **33**
2.1 Introduction . . . 33
2.1.1 Nativization . . . 34
2.1.2 Definitions . . . 35

2.2 Phonology . . . 36
2.2.1 Vowels . . . 36
2.2.2 Consonants . . . 37
2.3 Prosody . . . 39
2.4 Lexis . . . 40
2.5 Syntax . . . 42
2.6 Discourse and Pragmatic Features . . . 45
2.7 Discussion . . . 45
2.8 Conclusion . . . 49

3 The Chinese English Vowel Space Across Three Provinces 51
3.1 Introduction . . . 51
3.2 Methods . . . 54
3.3 Results . . . 56
3.4 Discussion . . . 62
3.5 Conclusion . . . 64

4 Durational Variation across inner/outer/expanding Circle Varieties of English 65
4.1 Introduction . . . 65
4.1.1 Durational variation in speech . . . 66
4.1.1.1 Durational Variation Metrics . . . 67
4.1.1.2 Other models of prosodic speech timing . . . 69
4.1.2 Acquisition of durational variation by non-native speakers of English . . . 69
4.1.3 Varieties under investigation . . . 70
4.1.4 The current study . . . 71
4.2 Methods . . . 72
4.2.1 Data . . . 72
4.2.2 Analysis . . . 73
4.3 Results . . . 73
4.4 Discussion . . . 77
4.5 Conclusion . . . 79
4.6 Appendix . . . 80

5 Towards a Vowel Formant Based Quality Metric for Text-to-Speech Systems 81
5.1 Introduction . . . 81
5.2 Method . . . 83
5.2.1 Dataset . . . 83
5.2.2 Models . . . 84
5.2.3 Vowel Formant Based TTS Quality Metric . . . 85
5.3 Results . . . 85
5.4 Discussion . . . 86
5.5 Conclusion . . . 89

6 Competent but Cold: Learning with a Pedagogical Agent Speaking Non-Native Varieties 91
6.1 Introduction . . . 91
6.2 Background . . . 93
6.2.1 Personalization principle . . . 94
6.2.2 Voice principle . . . 94
6.2.3 Credibility and intentionality . . . 96
6.3 Methods . . . 97
6.3.1 participants . . . 97
6.3.2 Materials . . . 99
6.3.3 Procedure . . . 100
6.3.4 Analysis . . . 101
6.4 Results . . . 101
6.4.1 Learning outcomes . . . 101
6.4.2 Characteristics . . . 102
6.5 Discussion . . . 103
6.6 Conclusion . . . 106
6.7 Acknowledgements . . . 106
6.8 Appendix . . . 107
6.8.1 Supplementary tables . . . 107
6.8.2 Survey . . . 108

7 Conclusion 113
7.1 Summary of Results . . . 113
7.2 Filling the Research Space . . . 116
7.3 Limitations . . . 117
7.4 Future Research . . . 118

8 Referencess 121

A Deutsche Zusammenfassung der Dissertation 153
A.1 Einleitung . . . 153
A.2 Zentrale Begriffe . . . 154
A.3 Ergebnisse . . . 155
A.4 Diskussion . . . 156
A.5 Schlussbetrachtung . . . 157

B Eidesstattliche Erklärung zur Eigenständigkeit 159

List of Figures

1.1 Kachru's Three Concentric Circles Model 11
1.2 China linguistic map 13
1.3 Standard Chinese monophthong inventory 14
1.4 Example of a complete transcription in Praat 19
1.5 Example of a mel spectrogram with speech labels 24
1.6 Proposed Chinese English vowel inventory 27

3.1 Standard Chinese monophthong inventory 52
3.2 Vowel spaces of female and male participants from Guangdong, Guangxi, and Yunnan 56
3.3 Vowel Spaces of participants from Guangdong, Guangxi, Yunnan and groups combined 57
3.4 Proposed, tentative Chinese English vowel inventory 62

4.1 Visual comparison of reading and interview data based on nPVI-V and rPVI-C 76
4.2 Visual comparison of reading and interview data based on VarcoV and VarcoC 76

5.1 Distribution of monophthongs in training and validation set 84
5.2 Occurrences of monophthongs in the validation set 84
5.3 LJ Speech Vowel Space, plotted as F1 - F2 space 86
5.4 LJ Speech Phoneme Space All Data Points 87
5.5 LJ Speech Phoneme Space Central 75% of All Data Points 87

6.1 Responses to linguistic background questions. 98
6.2 Responsesn to tech-savviness questions. 98
6.3 Proportion of correct and wrong answers on the speaker's origin per variety. 100
6.4 Overview of the responses to the multiple choice and open questions per lecture variety. 102
6.5 Voice characteristics rated by variety of PA 104

7.1 Chinese English vowel space with tense and lax vowel areas for the combined data set 114
7.2 Proposed, tentative Chinese English vowel inventory 115

List of Tables

2.1 Pronunciation of Vowels in Chinese English Keywords 46
2.2 Categorization of phonological features of Chinese English 47
2.3 Category A features appearing in both HKE and CollSgE 48

3.1 Overlap percentage of tense and lax vowel space per province and gender 57
3.2 Results for vowel formants from Kurskal-Wallis tests. 58
3.3 Results of the vowel area overlap analysis 59
3.4 Results of the support vector machine (SVM) classification analysis 60
3.5 κ interpretation table 60
3.6 Combined indicators of mergers (overlap analysis / SVM analysis) 61

4.1 Values of all DVMs of the varieties (reading) 74
4.2 Values of all DVMs of the varieties (interview) 74
4.3 Results for the influence of data elicitation method of DVMs from Kurskal-Wallis tests. 75
4.4 Comparison of nPVI-V values with the literature 77
4.5 Articulation rate and phonation rate of read data 78
4.6 Articulation rate and phonation rate of interview data 78
4.7 Results of Spearman's rank correlation ϱ of DVM metric results and articulation rate 80

5.1 Vowel Space Overlap 86
5.2 Phoneme Space Overlap 88

6.1 Results of Wilcoxon signed-rank tests comparing characteristic ratings. 103
6.2 Results of Wilcoxon signed-rank tests on characteristic ratings per variety. 103
6.3 Comparison of readability measurements of the presented lecture's text. 107
6.4 Chi-square test results on answers to the multiple-choice questions per language group 107
6.5 Chi-square test results on answers to the multiple-choice questions per variety 107
6.6 Chi-square test results on answers to the vocabulary test per language group 108
6.7 Results of Wilcoxon signed-rank test comparing characteristic ratings for the Italian data. 108

7.1 Excerpt of the categorization of phonological features of Chinese English 113
7.2 Combined indicators of mergers (overlap analysis / SVM analysis) 114

List of Abbreviations

AI	artificial intelligence
AmE	American English
ASCII	American Standard Code for Information Interchange
BrE	British English
ChE	Chinese English
CPA	conversational pedagogical agent
CSV	comma-separated values
CzE	Czech English
DVM	durational variation metric
EDTs	embodied digital technologies
GDPR	General Data Protection Regulation
HCI	human-computer interaction
IPA	International Phonetic Alphabet
IrE	Irish English
ItE	Italian English
L1	first/native language
L2	second/foreign language
MFA	Montreal Forced Aligner
NiE	Nigerian English
PA	pedagogical agent
PIPL	Personal Information Protection Law
ScE	Scottish English
SLA	Second Language Acquisition
SVM	support vector machine
TTS	text-to-speech

Acknowledgements

I would like to thank Josef Schmied, not only for supervising this work or encouraging me to get back into academia in the first place but for being a true mentor since the very beginning of my academic pursuits. Over several years and projects, he always encouraged and guided me to academic and personal growth.

I would also like to thank Christina Sanchez-Stockhammer for her useful evaluation of this thesis and for the opportunity to present my work to her students.

I also thank Chang Chenguang for the helpful thesis evaluation and his support with the data collection process and the evaluation of the text-to-speech system. Without his support, much of the work presented in this thesis would not have been possible.

I would like to thank the English linguistics research team, in particular Marina Ivanova, for the support and excellent cooperation, not only on the work presented in this thesis but also across multiple other projects.

Moreover, I would like to thank my colleagues Rewa Tamboli and Stefan Taubert of the D03 subproject of the collaborative research center "Hybrid Societies" (SFB1410) at Chemnitz University of Technology. The interdisciplinary collaboration with them has been very valuable and inspiring. I am grateful for the opportunity to work in the collaborative research center funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project-ID 416228727), which allowed me to fully focus on my research and provided many stimuli on my academic progress.

I would like to thank Wang Huabin for recording the Chinese English data, and Adeiza Isiaka for supplying part of the Nigerian English data. Furthermore. I would like to thank our student assistants Michelle Klein, Christin Reuter, and Natalia Przybysz for the countless hours spent on the transcription of our data.

I would also like to thank Robert Fuchs for the fruitful discussion on speech rhythm and outer and expanding circle varieties. Furthermore, I thank all anonymous reviewers, many of whom provided extensive feedback and suggestions from which I learned much.

Finally, I would like to thank my family and friends for their enduring support and encouragement to pursue my goals. Their kind words of encouragement got me through the more challenging phases of writing this dissertation.

1 Introduction

Both language and learning are at the core of human life. While researchers and philosophers do not agree whether reality shapes language or language shapes reality (cf. the Sapir/Whorf Hypothesis Hoijer, 1971), being able to communicate using spoken (and written) language has been a crucial factor in human development and evolutionary success (Prinz, 2012). Learning is one of the many things that tremendously benefits from the human ability to use language. The ability to efficiently share and transmit knowledge and lived experience is central to human life, in both spoken and written form (cf. functionalism and the Prague School Luelsdorf, 1983).

With the accelerating digital transformation of society (Hilbert, 2020) and the integration of embodied digital technologies (EDTs Mandl et al., 2023) into daily life (Rawassizadeh et al., 2019) and academia (Du Toit & Verhoef, 2018), language becomes even more important as a mode of interaction with these technologies (Marge et al., 2022) to ensure interaction between humans end EDTs to be as smooth as seamless as interaction between humans in these hybrid societies (S. Meyer, Mandl, Gesmann-Nuissl, & Strobel, 2022).

Especially in the context of teaching and learning, the potential of digital technology has often been praised, starting with asynchronous learning (Jaffee, 1997), multimedia learning (Mayer, 2014a) and lately also using robots (cf. Robot-Assisted Langage Learning Engwall, Lopes, & Åhlund, 2021), chatbots and large language models (Jeon & Lee, 2023). In the sub-project D03 "Credibility Through Non-native Language Varieties in Conversational Pedagogical Agents" of the collaborative research center "Hybrid Societies: Humans interacting with embodied digital technologies" linguistic variables relevant to a potential digital teacher, a conversational pedagogical agent (CPA) are investigated. The project envisions a conversational pedagogical agent (CPA) to be useful in delivering online classes and tutorials to foreign students at German universities in a hybrid, English medium learning environment

The project hypothesizes, based on previous research on German varieties like Austrian German and Standard German (Rey & Steib, 2013), that delivering content to students in their variety improves learning outcomes by improving the credibility of the CPA. Furthermore, it is hypothesized that some linguistic features contribute more to the attributed credibility than others. Therefore, it is necessary to identify these linguistic features. Lastly, since Chinese students are the largest group of foreign students at German universities, the project focuses on Chinese English (ChE) in particular.

To achieve the goal of the project, an interdisciplinary team of linguists, computer scientists, and psychologists developed a three-stage research process. In the first stage, relevant features of Chinese English were identified and analyzed by the linguists (see chapter 2, chapter 3, and chapter 4). In the second stage, a text-to-speech (TTS) system, based on Tacotron2 (Shen et al., 2018) a state-of-the-art neural speech synthesis system, was built. The system was further extended to accept input text in the Internationa Phonetic Alphabet (IPA) rather than standard English orthography. This allows modifying phonemes either automatically, based on rules and probabilities, or manually, before synthesis. Prosodic cues can also partly be controlled using annotation of the input text before synthesis, by adding primary, secondary, or no stress to syllables nuclei. Similarly, it is possible to add pauses of different lengths either between words or within words. Other prosodic cues are learned implicitly. In the third stage,

the output of the TTS system was evaluated and learning studies were conducted to test whether the linguistic features implemented improved the learning outcome over a baseline of American English produced by the same TTS system.

1.1 Embodied Digital Technologies and Artificial Agents

In an increasingly digitized world, human-computer interaction (HCI) has evolved dramatically. One of the most profound and transformative shifts in this landscape is the emergence of embodied digital technologies. These technologies have transcended traditional boundaries, creating a synergy between the physical and digital realms (Pustejovsky & Krishnaswamy, 2021). Through the integration of sensory perception and physical presence, they have redefined how we interact with and perceive the digital world, offering unprecedented opportunities and challenges.

The idea of EDTs interacting with the physical world, traditionally imagined as robots[1], gained traction from the middle of the 20th century onwards, even though the general idea of artificial agents, or *automata*, can be traced back to Ancient China (Needham, 2005) and Ancient Greece (Gera, 2003). The developments of robots in the 20th century coincided with the rise of artificial intelligence (AI) starting with the seminal paper by Turing (1950). In combination with advances in engineering, this led to the development of a myriad of EDTs, ranging from manufacturing robots and social robots (e.g. Breazeal, 2003; Tavani, 2018) to autonomous cars and virtual agents integrated into smartphones.

A key to successful interaction with embodied digital technologies and artificial agents is communication. As language is crucial in communication, these agents' voices must be credible in order for them to be accepted as discourse partners by humans (Schmied, 2020).

1.2 Intentionality and Credibility

1.2.1 Intentionality

Intentionality is a central and enduring concept in philosophy that has been explored and developed by philosophers over millennia. It refers to the inherent property of mental states and acts to be "about" something, involving a directedness or reference to objects or states of affairs in the world.

The earliest, indirect accounts of intentionality can be found in Plato's theory of Forms. His conceptualizations of thoughts and knowledge always include a directedness towards the Forms, as they are immaterial and have to be about something, making cognitive processes inherently intentional (for a detailed discussion on thoughts and knowledge in Plato's work, see Cornford, 2013; N. P. White, 1987). Aristotle affirms the directedness of thinking and knowledge towards the universal essence of things, away from specific instances. He further extends intentionality to perception and introduces *phantasia* (for a detailed account of the meaning of *phantasia* see K. White, 1985) as an intermediary between instances and essence (Caston, 1998).

The stoics further develop the concept of *phantasia*, explicitly adding the dimension of representation, which is a pervasive feature of stoic thought. For the stoics, all mental states are either

[1] The term entered popular consciousness through the works of the Czech brothers Joseph and Karel Capek (1925; 1920), who derived it from the Czech word *robota*, meaning serf or laborer(Hockstein, Gourin, Faust, & Terris, 2007). Isaac Asimov is generally credited with popularizing robotics through the "Three Laws of Robotics" presented in his short stories (Asimov, 1950: 40).

representations or assent to a representation (Caston, in press: 3) as they analyze all mental states in terms of representation. Therefore, stoic psychology is representational, or intentional.

The idea of representation is revived by Franz Brentano, whose work serves as an entry point to modern discussions on intentionality[2]. Brentano considered intentionality to be the directedness towards an object[3] that might only exist in the mind. Brentano's central thesis (Brentano, 1874) is that intentionality is "the mark of the mental" (Crane, 1998: 229)

Brentano's student Edmund Husserl agreed with the directedness towards an object (which he called *noema*), however, he disagreed with the idea that intentionality is a purely mental phenomenon and added a transcendent dimension to intentionality (Christensen, Welch, & Barr, 2017: 116). Intentional objects, in fact all objects of consciousness, are inherently transcendent. As such, intentionality became central to the phenomenological method developed by Husserl (Moran, 2000: 41).

At this point, a major split happened in the history of Western philosophy: Husserl and the followers of his phenomenology established the Continental tradition of philosophy. While Heidegger initially kept Husserl's central concept of intentionality and criticizes it as being limited by a Cartesian standpoint (cf. Moran, 1996: 28), he later (especially in Heidegger, [1927] 2006) abandons intentionality altogether, replacing it with his central, existential concept of being-in-the-world. Merleau-Ponty shifts intentionality away from the transcendental by emphasizing its embodied nature (Halák, 2023; Kelly, 2016; Reuter, 1999). Ultimately, Derrida, with his deconstruction movement, discards intentionality altogether (Alvis, 2016; Moran, 1996; Vandevelde, 2005).

The Analytic (North American) tradition of philosophy retains intentionality as a central concept. Chisholm (e.g. 1990) introduced the ideas of Brentano and Husserl to the North American audience by drawing attention to the logical features of intentional idioms (Moran, 1996: 29). Dennett (2010) attempts to dissolve the general Cartesian dualism that imbued the general debate on intentionality by embracing naturalism at a micro level while maintaining intentionality at the macro level for its explanatory power of human rationality (Moran, 1996: 30). Similarly, Searle (1983) tries to naturalize intentionality by adopting a view of "biological naturalism" (Moran, 1996: 30), positioning that while the mental is caused by physical processes in the brain, it still is a genuine, higher-order property of the brain (Searle, 1992). Notably, Dreyfus (1972); Dreyfus and Dreyfus (1986) maintains Heidegger's transcendental view of intentionality as being outside the physical as part of his general critique of artificial intelligence[4].

In a recent contribution to the discussion of intentionality (and ultimately consciousness), Prinz (2012) develops the interesting idea of social mirroring to the debate. While he maintains the view that intentionality results from physical processes in the brain and thus is not transcendental, he proposes that instead of projecting our innate consciousness outwards, we import consciousness through experiencing the intentional actions of others in a process that he calls social mirroring.

In a linguistic sense, credibility is often seen as related to the first definition of intentionality, in the sense of "intent toward receiver", which includes several dimensions, such as "competence", "trustworthiness", and "goodwill" (McCroskey & Teven, 1999: 90). A linguistically credible speaker therefore

---

[2] Brentano did, however, build on the discussion of the Scholastics during the Middle Ages (see Brentano, [1874] 1973: 124) For a detailed account of Brentano's reception of the Scholastics, see Brentano and Hedwig (1980).

[3] Brentano uses the term object in the classic Aristotelian and scholastic sense of formal object, which is the aspect according to which a material object becomes accessible to the mind. However, mental objects need not correspond to material objects (Antonelli, 2022: 310).

[4] Interestingly, while Searle disagrees on the transcendental nature of intentionality, both agree in their criticism of AI.

expresses goodwill implicitly through language cues. This goodwill can also be considered a form of caring for others, which has been shown to increase teachers' credibility in human educational contexts (Teven, 2007: 435).

Consequently, the linguistic credibility attributed to CPAs could be improved by imitating users and integrating their feedback (Maes, 1994: 40) and preferences, to avoid negative face threat (P. Brown & Levinson, 1987). At the same time, evoking stereotypes by mirroring non-native varieties (cf. *Chinglish* in chapter 2) might lead to negative face threat (Labov, 1994: 78) and thus should be avoided. Overall, creating linguistically credible CPAs that are suitable for individualization and context-aware has the potential to increase their acceptance and reduce negative stereotypes (Sutton, Foulkes, Kirk, & Lawson, 2019).

1.2.2 Credibility

The attribution of credibility to CPAs, and EDTs in general, is contingent on intentionality. There are two competing and often conflated definitions of intentionality: the psychological definition (E. Craig, 1996; Honderich, 1995; Mele, 2009), which is close to the colloquial use of intention, in which an actor commits to an action, and the philosophical definition, which relates intentionality to having intentional states, i.e. mental representations of reality (Brentano, 1874; Prinz, 2012; Searle, 1983). While the possession of intentionality seems to be a uniquely human (Searle, 1980) that is not accessible to EDTs (though AI advocates might disagree[5]).

Following the theory of social mirroring (Prinz, 2012, 2013), it is irrelevant whether EDTs have intentional states, as intentionality is attributed to them by humans they interact with (cf. anthropomorphism Duffy, 2003; Epley, Waytz, & Cacioppo, 2007). Prinz (2012) proposes an approach to a philosophy of mind that is based on import, rather than export. This theory departs from the naturalistic view that self-consciousness is an innate function of humans and thus forms the basis of all mental functions (e.g. Descartes, 1641; Kant, 1781) and adopts a more constructivist approach that posits that humans construct their self-consciousness by mirroring, or importing, what they see in others into themselves. Consequently, consciousness and intentionality are no longer private phenomena but minds are open to (and dependent on) other minds.

This *Open Minds* approach is not limited to humans but can integrate other entities. Prinz admits animals have a similar position to humans in an open minds system but does not take a stance on artificial agents (Prinz, 2012: 205 footnote 15). However, irrespective of artificial agents' ability to construct consciousness, due to anthropomorphic processes and the import-based nature of the open minds approach, they can easily be integrated into the construction of humans' minds. Thus, the conversational pedagogical agent envisioned in this research project becomes an actor in a system of minds and therefore a co-constructor of the minds of its users. This ultimately allows a CPA to take a

[5] Even though Searle's original article already addressed several possible replies that were prominent in the academic discourse at the time (Searle, 1980: 419-422) and was published with open peer comments (e.g. Abelson, 1980; Block, 1980; Bridgeman, 1980; Danto, 1980; Dennett, 1980; Eccles, 1980; Fodor, 1980; Haugeland, 1980), it attracted criticism for a much longer period (e.g. Churchland & Churchland, 1990; Cole, 1991; Copeland, 1993; Dennett, 2013). Despite all criticism, the argument still stands and there is no final consensus on the limitations of computational accounts of mind. In the four decades since, computers have become more powerful, as has software, however, it does not seem that general, strong AI is any closer to becoming a reality than in 1980, irrespective of attempts by marketing departments of technology companies to convince customers otherwise. Consequently, I agree with Searle on the limits of computers being able to simulate minds.

similar role as a human teacher (not necessarily for the same tasks or purposes) and requires it to be equally credible in order to be accepted as a discourse partner (Schmied, 2020).

Credibility is a multifaceted phenomenon, typically associated with "competence", "trustworthiness" (Kohring, 2001), and "goodwill" (McCroskey & Teven, 1999).To maximize the usefulness of a CPA, it would be desirable if users were to attribute goodwill and credibility. In a non-native English medium educational setting, a CPA might evoke these positive attributions by speaking in the same variety as its users, their in-group variety (for a review of work on in- vs. out-group preferences see Molenberghs, 2013), facilitating social mirroring effects. Thus, linguistic credibility becomes an expression of goodwill, which has been shown to improve learning outcomes (Finn et al., 2009b; Teven, 2007). Leveraging this effect to improve the credibility of an artificial agent is especially desirable in an educational setting in which the agent takes the role of a knowledge provider, as humans are generally attributed greater source credibility than artificial agents ("human superiority effect", see Finkel & Krämer, 2022). Source credibility is an important aspect of accepting information, especially from technology, including websites (Metzger & Flanagin, 2013), smart speakers (Cho, Sundar, Abdullah, & Motalebi, 2020), or social robots (Edwards, Edwards, Spence, Harris, & Gambino, 2016). The attribution of credibility to (embodied) digital technologies can also occur at different stages in the human-machine interaction process. Fogg and Tseng (1999) differentiate presumed, reputed, surface, and experienced credibility. Presumed and reputed credibility ascriptions happen before the encounter with the digital technology, surface credibility is based on first impressions, and experienced credibility becomes relevant after an interaction or in repeated interactions.

In conclusion, increasing the credibility attributed to a CPA is highly desirable to mitigate the "human superiority effect" and to establish the agent as a trustworthy knowledge provider that acts with goodwill towards its users. In the research project that this thesis is based on, the agent's voice is designed to match the users' non-native variety of English to increase its credibility and thus enhance learning outcomes.

1.3 Sociolinguistic Theory

Around the time of the inception of the field of sociolinguistics, the field of linguistics was dominated by the conflict around Chomskyan transformational and generative grammar (Chomsky, 1957, 1965). While Chomsky and his followers solely focused on linguistic *competence* exemplified by idealized speaker-hearers and dismissed language in use, or *performance* in their terminology, as irrelevant (Wodak, Johnstone, & Kerswill, 2010). However, some scholars (e.g. Habermas, 1981; Hymes, 1968, 1974) resisted the dismissal of language in use as irrelevant and formulated theories of *language as action* and *communicative competence*, following John Searle's *Speech Act Theory* (Searle, 1969). At its core, this conflict can be traced back to Saussure and the European neogrammarians of the late 19th and early 20th century, who, based on the axiom of categoricity (see Chambers, 2009: 12), considered the social uses of language outside the realm of science (see Schuchardt 1885 for contemporary criticism, or Joseph 2004 for more recent criticism of the decoupling of language from its use).

1.3.1 Variationist Sociolinguistics

The field of sociolinguistics with its many sub-disciplines has undergone many changes and developments since its inception at the Linguistics Institute at Bloomington in 1964 (Spolsky, 2010: 11). In the

following section, I want to highlight some of these developments to illustrate the logical connection with research in the field of World Englishes presented in section 1.3.2. The term sociolinguistics was first used in English by Cuerie (1952), however Uriel Weinreich, whose 1953 paper on language contact is still frequently cited to this day, is frequently credited with establishing the term in academic discourse. Weinreich's most prolific and influential student William Labov added the variationist dimension to sociolinguistics. Based on the methods of quantifying language data by Fischer (1958), Labov set out to "describe correlations between patterns of linguistic behaviour and social identities in as objective a way as possible" (Hazen, 2010: 26). By focusing on phonetic variation and change in his two seminal works on Martha's Vinyard (1963) and New York City Department Stores (1966), Labov demonstrated that sound change was both synchronically observable and tied to social factors in a community (Hazen, 2010: 32). Another central concept pioneered by Labov is that of the linguistic variable, a term first used in his 1963 publication, denoting "a variable element within the system controlled by a single rule" (Weinreich, Labov, & Herzog, 1968: 159). Through his work, Labov shaped the field of sociolinguistics like few others, both theoretically and methodically. In terms of linguistic theory, he postulated that language is heterogeneously ordered, making language change regular, and he proposed his principles of linguistic change (see Labov, 1994, 2001, 2010) that continue to influence variation sociolinguistics today. While these theoretical principles seem to be a departure from the established ways of studying language at the time, their origins can be traced back to the European neogrammarians (see J. Milroy & Milroy, 1985; Murray, 1994), whose ideas had been popularized in the US by Edward Sapir (e.g. Sapir, 1929) the American Structuralists of the early 20th century (cf. L. Milroy, 1997). Even though Labov distanced himself from the American Structuralists (Labov, 1972), his methods of analyzing phonetic change can be traced back to Hockett (1958), a structuralist. His focus on accountability of data (L. Milroy, 1997: 78) further connects him with the structuralists. Labov's theoretical contributions to the field of sociolinguistics might have lost their traction over the years, but his methods, which the variationist sociolinguistic tradition is based upon, however, are still relevant today (e.g. Tagliamonte, 2012).

1.3.1.1 The Research Framework

The variationist sociolinguistic research framework is based on the vernacular, the sociolinguistic variable, the principle of accountability, and the circumscription of the variable context (Tagliamonte, 2012). The focus on the vernacular implies collecting data in a setting where subjects are least likely to control their speech (Labov, 1984). The linguistic variable has to have at least two settings with similar meanings (Tagliamonte, 2012: 2) should be relatively unconscious to the subjects (Tagliamonte, 2013: 383). The principle of accountability (e.g. Labov, 1966: 49) dictates that all possible variants have to be considered in the analysis, not just the interesting one, essentially preventing cherry-picking data to fit a hypothesis. The circumscription of the variable context[6] is the sometimes lengthy process of identifying and describing all variants of a sociolinguistic variable that are functionally parallel (Tagliamonte, 2012: 10) and can vary.

This theoretical basis is then operationalized by applying quantitative and qualitative methods (Bortz & Döring, 2006). The quantitative component of the analysis is correlating the dependent variable, i.e. linguistic variation, with one or more independent variables, i.e. social or linguistic categories. (Chambers, 2009: 18). This quantitative, probabilistic approach sets the variationist sociolinguistic

[6] This process is also called identifying the "envelop of variation" (L. Milroy & Gordon, 2003: 180).

research framework apart from categorical models (employed by Chomsky for example). The correlation analysis is usually done via regression modeling, which also allows ranking the effect of dependent variables, which is important for the qualitative part of the framework, i.e. describing temporal, social and linguistic patterns of language variation and change (Tagliamonte, 2013: 382). In summary, the variationist sociolinguist research framework can be conceptualized as follows (Tagliamonte, 2012: 7):

1. Observation – hear and/or see variation in language use;
2. Identification – select the linguistic variable for study;
3. Reconnaissance – determine if the variation occurs and where;
4. Systematic Exploratory Observation:
    - What is the inventory of forms?
    - What are the patterns?
    - When does the variation occur and under what circumstances?
    - Who uses the variation and how?
5. Test hypotheses, claims, and observations;
6. Interpret and explain the variable patterns, social and linguistic.

The last two items usually entail statistical modeling, which enables the research to answer the following questions (Tagliamonte, 2012: 7):

- Which factors are statistically significant?
- What is the relative contribution of the independent factors tested in the model?
- What is the order (from more to less) of factors within the independent factors (predictors), the constraints or constraint hierarchy?
- Does this order reflect the direction predicted by one or the other of the hypotheses being tested?

In conclusion, the variationist sociolinguistic research framework combines quantitative and qualitative methods to produce reliable and replicable results.

1.3.1.2 Common Variables

While the variationist sociolinguistic research framework is suitable for investigating a variety of linguistic and social categories, several categories have received intense attention over the last decades: social class, sex, age, style and register, and ethnicity and culture.

One of the earliest studies in the field, Labov's famous New York Department Store study (Labov, 1966), investigated the relationship between variation in speech and social class. The term class is used descriptively here, to model the socioeconomic hierarchy of a community, and not prescriptively to fit people in a set of identifiable classes (Labov, 2001: 113). From the earliest days of sociolinguistic research, studies have shown meaningful correlations between linguistic variables and class, evidenced by systematically ordered (i.e. stratified) use of variants by different portions of the community depending on social status (Labov, 1972: 8). While Labov (2001: 114) argues that there have to be at least four social strata to generate meaningful results, numerous subsequent studies have shown that the linguistic variable reflects the social structure even for binary contrasts (Chambers, 2009).

Sex[7] is the linguistic variable that in most studies consistently produced the clearest contrasts: women tend to avoid stigmatized (i.e. associated with low social status) forms (Tagliamonte, 2012: 32). Explanations of this effect range from biology (Chambers, 2009: 145-151) to social norms and practices (Eckert, 2000: 196), including access to the linguistic market (cf. Sankoff & Laberge, 1978). Similarly to social class, sex as a linguistic variable and the numerous studies showing clear male—female patterns in language use are oversimplifications of social realities. Consequently, many studies have focused on individual variation within these groups in an attempt to explain the observable patterns (Eckert, 1988, 2000; Eckert & McConnell-Ginet, 1999).

While it is intuitively apparent that language correlates with the age of speakers, differentiating whether variation observed in an apparent time study represents language change in progress or age-graded speaker behavior is difficult (Cheshire, 2005: 1553). While age-grading speakers is problematic because it is always arbitrary (Llamas, 2007: 71) and interacts with culture, based on social or religious rituals (Cheshire, 2004: 1552), Chambers (2009) argues that it is realized predictably, keeping the hypothesis that individuals are linguistically stable despite aging (Chambers, 2009; Labov, 1994) intact.

Style refers to the widely attested pattern that people tend to use high prestige variants in more formal styles of interaction, while register denotes individual performance in a certain social setting (Tagliamonte, 2012: 34-35). While there is certainly more to style than the formal—informal continuum, especially if speakers are granted an active role in shaping their language (Schilling-Estes, 1998), isolating it is difficult, as it interacts with sex and social class (Tagliamonte, 2012: 23).

Ethnicity and culture in sociolinguistic research often focus on differences in language use between the dominant or mainstream community and the ethnic community (Tagliamonte, 2012: 38) that can act as social barriers (Trudgill, 2000: 45-46). Studies in the North America have focused on different ethnic groups, such as African Americans (e.g. Wolfram, 2000), Hispanics (e.g. Santa Ana, 1996), Italians or Chinese (e.g. Hoffman & Walker, 2010). An interesting aspect of the majority—minority dynamic is the possibility of ethnic minority speakers driving change in the whole speech community (Horvath, 1985).

1.3.2 World Englishes

The origin of the term *World Englishes*[8] is not exactly clear, however, Kachru (1984, 1985) is often credited with coining the term after laying the groundwork for the concept in his earlier work in Indian English (Kachru, 1965). Possibly it simply arose organically in the academic discussions of the status of the English language in post-colonial countries. The plural form *Englishes* might seem unusual at first, yet it emphasizes the plurality and diversity found in the English language today and highlights that there is no single normative center (Mesthrie & Bhatt, 2008: 3).

It is certain, however, that the field of World Englishes is rooted in the interactions of the English language with local languages and cultures as the English language expanded with the borders of the British Empire (D. Crystal, 2003). The stages of this expansion can be traced back several centuries and are usually separated into four diasporas that roughly correspond to the historical timeline (see Nelson,

[7] The debate on biological sex and socially constructed gender has featured prominently in sociolinguistic research (cf. Cheshire, 2004: 423-427), with some scholars arguing keeping sex as the major sociolinguistic variable for the sake of replication of studies (Labov, 1990: 209), while others criticized the generalizations made as lacking empirical foundation (Cheshire, 2004).

[8] *New Englishes* is a similar term frequently used in the literature. Both terms seem to be used interchangeably.

Proshina, Davis, Kachru, and Kachru (2019) for a detailed discussion): (1) The first diaspora marks the spread of English throughout the British Isles in the Middle Ages. (2) The second diaspora refers to English spreading to the New World, meaning North America, Australia and New Zealand but also the Caribbean. (3) The third diaspora usually denotes the spread of English throughout much of Africa and Asia. (4) The fourth diaspora includes the spread of English in South America, Europe, East Asia, and China. While the borders of the Empire shrunk again, especially through independence movements and the accelerating process of decolonization (Betts, 2006) since the middle of the 21st century, the English language often continued to play a crucial role in these postcolonial societies, either as an official language, a medium of instruction, or a lingua franca among diverse linguistic communities.

As a field of study, World Englishes seeks to describe and understand varieties of English spoken and used in various sociolinguistic and cultural contexts around the world. This makes it rather similar to the variationist sociolinguistic research framework discussed in section 1.3.1.1, in both theoretical approach and methodological operationalization. Variation is a fundamental aspect of World Englishes research as it explores how English has evolved and adapted within different linguistic communities, leading to the emergence of unique regional and national varieties. Some of the aspects researched under the World Englishes research paradigm sound familiar from the discussion of sociolinguistic variables: (1) social factors (2) regional variation (3) identity and prestige (4) language contact. At the core of both areas of research lies the assumption that people's choice between two or more linguistic variants is due to external, social factors.

Both variationist sociolinguistics and World Englishes consider social factors as essential in understanding language variation. Variationist sociolinguists have long recognized the influence of social variables on linguistic choices within specific communities. In World Englishes research, these social factors extend to the global scale, examining how sociopolitical, cultural, and historical contexts shape the development of English varieties in different regions (Mesthrie & Bhatt, 2008: 1, 92-96).

While variationist sociolinguistics often focuses on regional dialects within a single language community, World Englishes extend this concept to encompass regional varieties of English spoken around the world. Schneider's Dynamic Model of Postcolonial Englishes, for example, considers how regional factors influence the development of English in postcolonial settings (E. W. Schneider, 2007). The model will be discussed in greater detail in section 1.3.2.2.

Both fields recognize the importance of linguistic identity and prestige in shaping language variation. In variationist sociolinguistics, linguistic choices are often associated with identity and social status within a particular community. In World Englishes, identity and prestige are explored on a broader scale, considering how speakers of non-native English varieties navigate issues of identity, prestige, and linguistic discrimination in global contexts (E. W. Schneider, 2014a).

Language contact is a central theme in both fields. Variationist sociolinguistics may examine the effects of language contact within a community, while World Englishes researchers investigate how English interacts with other languages in multilingual settings, leading to phenomena such as code-switching and borrowing (Mesthrie & Bhatt, 2008: 156-199).

Even though World Englishes studies the effects of language contact between English and local languages, it does not include contact languages such as pidgins and creols[9]. World Englishes might be a separate phenomenon, but research on pidgins and creoles can still "illuminate the study of New Englishes" (Mesthrie & Bhatt, 2008: 198)

[9] For an accessible introduction to the topic of (English) pidgins and creoles, see Holm (2000).

To categorize the multitude of World Englishes, several models have been proposed over the last decades. These models generally fit into two major categories: static models and dynamic or evolutionary models (E. W. Schneider, 2014a: 36-37). The earliest static models were nation-based and focused on standard English(es). Examples include Bailey and Görlach (1986); Trudgill and Hannah (2013). The most commonly known types of static models are tripartite models, including the models by Strang (1970) and, most famously, Kachru (1985). Other types of static models include hub and spokes models (e.g. Görlach, 1990; McArthur, 1987) and listings (e.g. Gupta, 1997; McArthur, 1998). The two most prominent dynamic models are Trudgill's Deterministic Model (Trudgill, 2004) and Schneider's Dynamic Model (E. W. Schneider, 2007). Both Kachru's model and Schneider's model have gained significant traction in the research community, whereas Trudgill's model has generated the most controversy (see Trudgill (2008) and responses).

1.3.2.1 Kachru's Concentric Three Circles

Kachru's model of concentric circles (1985) is a seminal framework for understanding the global spread and diversification of the English language. The model categorizes the varieties of English into three concentric circles, each representing a distinct sociolinguistic context:

- Inner Circle: This circle comprises countries where English is the native language and holds the status of a mother tongue. The Inner Circle includes regions such as the United Kingdom, the United States, Canada, Australia, and New Zealand. These varieties are also called norm-providing.
- Outer Circle: In the Outer Circle, English functions as a second language. It is used for various domains like education, administration, and literature, and it often coexists with indigenous languages. Countries in the Outer Circle include India, Nigeria, Singapore, Malaysia, and South Africa. These varieties are also called norm-developing.
- Expanding Circle: The Expanding Circle consists of countries where English is primarily learned as a foreign language. In these nations, English serves as a tool for international communication, trade, and diplomacy. This circle encompasses many European, Asian, and South American countries. These varieties are also called norm-dependent.

Norm-dependent states may look to norm-providing states as linguistic models and sources of prestige, often adopting native-speaker norms in education and formal contexts. However, this relationship is not one-sided. Norm-providing states also adapt to the linguistic diversity found in norm-dependent states, as English varieties in these regions undergo their own development and standardization. The model, illustrated in figure 1.1 with the countries relevant to this thesis assigned to their respective circles, highlights the complex roles that English plays in different parts of the world, emphasizing its linguistic, cultural, and sociopolitical diversity. This framework has been influential in the study of World Englishes, providing a systematic way to categorize and analyze the various forms of English spoken globally.

1.3.2.2 Schneider's Dynamic Model

Schneider's Dynamic Model of Postcolonial Englishes (2007) is a comprehensive framework that explores the development and diversification of English in postcolonial contexts. This model provides

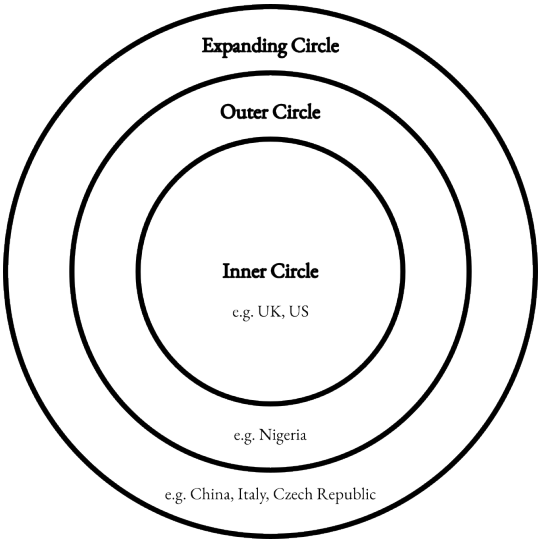


Figure 1.1: Kachru's Three Concentric Circles Model (1985)

insights into how English varieties emerge and evolve as a result of historical, sociopolitical, and linguistic factors. The model defines five phases in the process of the formation of a new variety of English:

1. Phase One – Foundation: In this phase, English is introduced through colonization. It serves as a tool of communication and administration but often coexists with indigenous languages. The linguistic features of English in this phase may resemble the colonial variety.

2. Phase Two – Exonormative Stabilization: English becomes more established as a result of formal education and administration. It is often influenced by prescriptive norms from the colonizing country. Varieties in this phase may exhibit more stability and standardization.

3. Phase Three – Nativization: English begins to undergo nativization, adapting to local linguistic and cultural norms. Local varieties with distinctive features emerge. Language contact with indigenous languages and cultural factors play a significant role in shaping these varieties.

4. Phase Four – Endonormative Stabilization: In this phase, the nativized variety undergoes stabilization and standardization, often with local norms and linguistic features. The variety gains prestige and is used in various domains, including literature and media.

5. Phase Five – Differentiation: English varieties continue to diversify within the postcolonial context. Distinct regional and national varieties develop, each with its own linguistic characteristics. These varieties coexist with the colonial or standard variety.

The model emphasizes the agency of local communities in shaping their English varieties and challenges the notion of English as a monolithic entity. It recognizes the dynamic nature of English in postcolonial settings and the importance of historical and sociopolitical factors in shaping these varieties.

1.3.2.3 On the Native Speaker

In the World Englishes framework, English is viewed as a dynamic and evolving global language spoken by diverse communities worldwide. This concept questions the hierarchy that has historically privileged native speakers and calls for a more inclusive and equitable approach to understanding and valuing English language users.

Traditionally, a native speaker of English is defined as someone who acquired the language as their first language from birth. Native speakers have been historically considered the benchmark for linguistic proficiency and authenticity. However, this definition fails to acknowledge the vast number of proficient and fluent non-native English speakers who use the language effectively and contribute significantly to its global use (D. Crystal, 2003: 147).

In World Englishes research, the focus shifts from a rigid emphasis on native speaker linguistic competence to a broader understanding of identity and sociolinguistic competence. Nonnative speakers often include their own linguistic and cultural practices in their use of the English language, which can manifest in their way of speaking (Jenkins, 2012) but also in everyday language acts (see Pennycook 2010 for examples). Language thus becomes a product of its use, removing the prescriptive constraints of the traditional native speaker model. Language norms develop through the interaction of language users. Ultimately, World Englishes research tries to progress from the dichotomy of *us* and *them* to a more inclusive concept of '*WE*-ness' (Kachru, 1992b: 2).

1.4 The Linguistic Landscape of China

To understand possible first/native language (L1) transfer effects[10] in second/foreign language (L2) English spoken by people from China, taking a closer look at the diverse linguistic landscape in China is necessary. At the macro level, there are languages of six major language families present in the PR China: Sino-Tibetan, Austro-Tai, Austronesian, Altaic, Austro-Asiatic and Indo-European (M. W. Lee 1987[11], cited in Tang and van Heuven 2009, see also Kurpaska 2010). At the meso level, there is a multitude of languages belonging to the different language families listed above. The majority of the languages commonly referred to as Chinese languages belong to the Sino-Tibetan language family. While Chinese dialectologists do not generally agree on how to classify the languages of the Sino-Tibetan language family, there is a general consensus that there are two major groups, the (Northern) Mandarin branch, and the Southern (non-Mandarin) branch, as illustrated by figure 1.2. In general, the level of mutual intelligibility is considered lower in the Southern branch (Yan, 2006). At the micro level, there are local languages, often referred to as "dialects", a translation of the Chinese term "方言 *fāngyán*" with the literal meaning "regional speech". However, the meanings of the Chinese and English terms differ in that the English "dialect" requires the parameter of mutual intelligibility (see Chambers & Trudgill, 1998), whereas the Chinese term *fāngyán* does not (Kurpaska, 2010). The Chinese term can refer to both languages, such as Mandarin, or a regional variety, such as Beijing dialect (Kurpaska, 2010: 2). While many of these "dialects" are not mutually intelligible, and might be "as diverse as the several Romance

[10] This presupposes one's preferred theory of Second Language Acquisition (SLA) allows for L1 interference (cf. Full Access Full Transfer Hypothesis (Schwartz & Sprouse, 1994, 1996; Sprouse & Schwartz, 1998), Full Access Hypothesis (Epstein, Flynn, & Martohardjono, 1996, 1998), Failed Functional Features Hypothesis (Hawkins & Chan, 1997)). A detailed discussion of second language acquisition theories would, however, be beyond the scope of this thesis.

[11] This classification is not undisputed, alternatives have been proposed by F.-K. Li (1973), Ramsey (1989), and Eberhard, Simons, Fennig, and Summer Institute of Linguistics (2022).

languages" (Chomsky, 1986: 15), or as English from Dutch (Chao, 1980: 96), the traditional English rendering of *fāngyán* as "dialect"[12] will be used here. While the political dimension of language use and terminology is beyond the scope of this thesis, it cannot be completely ignored, as there have been substantial efforts to promote the use of (Beijing) Mandarin, also known as *Putonghua*, the most widely spoken Chinese language commonly referred to as the standard form (Chao, 1980: 96), throughout Chinese society, especially in education (Kurpaska, 2010; C. Lin & Jackson, 2021).



Figure 1.2: "China linguistic map" U.S. Central Intelligence Agency, marked as public domain, Wikimedia

Chinese dialects vary considerably in every aspect of their pronunciation, the study of which usually focuses on the syllable as a unit of analysis. Standard Chinese syllables consist of an initial consonant, a final, and a tone (Duanmu, 2007: 48). The final consists of a required nucleus, which is a vowel, and optionally is preceded by a medial and/or followed by a coda, which again is a consonant. There is considerable variation in the syllable initial, with southern varieties generally having fewer syllable initial consonants (Norman, 1988: 212). Similarly, the syllable finals also vary considerably, both in the optional medial and coda and the vowel. The vowel systems of Chinese dialects range from only

[12] The question of whether a variety of spoken language is a "dialect" or a "language" also has a political dimension as well as a dimension of ethnic identity. However, a discussion of these dimensions would be well beyond the scope of this thesis.

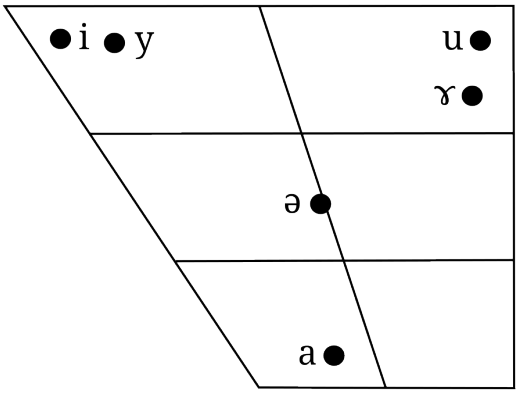


Figure 1.3: Standard Chinese monophthong inventory (from Zee & Lee, 2007: 110)

three vowels to eleven vowels (Zee & Lee, 2007). The six Standard Chinese monophthongs in their characteristic triangular shape are illustrated in figure 1.3. The last required element of a Chinese syllable, the tone, also varies with dialects. The number of tones ranges from three to seven, with northern dialects typically having fewer tones than southern dialects (Norman, 1988: 9). Standard Chinese differentiates four tones (W.-S. Lee & Zee, 2003), while Cantonese mostly maintains the eight Middle Chinese tones (Norman, 1988: 217-218).

1.5 Creating a Research Space

Building on the theoretical background described above, the research project *D03 Credibility Through Non-native Language Varieties in Conversational Pedagogical Agents* investigates the effect of credibility attributed to embodied digital technologies, in particular to conversational pedagogical agents, on learning outcomes of university students. Previous research on German varieties like Austrian German and Standard German (Rey & Steib, 2013) has shown that a human teacher speaking a local variety in contrast to the standard variety is perceived positively and even improves learning outcomes. This project extends this idea in two aspects: (1) instead of a first language (L1) educational context, a second/foreign language context is investigated; (2) instead of a human teacher an artificial conversational pedagogical agent will deliver the content. The goal of the research project is to investigate which specific non-native language cues increase the credibility attributed to a CPA by its users. The target users of the CPA are Chinese students at German universities, who in the future might use an e-learning system in a hybrid, English medium learning environment. However, other exchange students from the Czech Republic, Italy, or Nigeria for example could also be targeted, based on the data collected in the project.

While there is previous research on the credibility of robots (Finkel & Krämer, 2022), artificial teachers (K. N. Moreno, Klettke, Nibbaragandla, & Graesser, 2002) and teacher credibility (for a meta-review see Finn et al., 2009b), these strands of research have previously not been merged. This is what the research project D03 mentioned above and this thesis attempt. By extending the idea that a credible teacher improves the learning outcomes of students to an artificial teacher in the sense of a conversational pedagogical agent, the topic is integrated into the wider research area of *Hybrid Societies*.

To investigate the effect of such a CPA speaking in a non-native variety, a text-to-speech system that allows synthesizing non-native speech has to be developed and evaluated (cf. chapter 5). Since the research project focuses on the effect of linguistic credibility, a linguistically motivated quality metric for TTS systems would be desirable, yet has not been developed previously. Additionally, most existing metrics to assess the quality of synthesized speech focus on overall quality, whereas the metric presented here explicitly focuses on vowel quality, which presents a new approach.

As our target user group of a future CPA speaking non-native English is Chinese exchange students at German universities, systematizing previous work on Chinese English into a preliminary account of the variety similar to other Asian expanding circle varieties, such as Hong Kong English or Singapore English, is necessary (cf. chapter 2) and had not been done before. The systematic overview provides both an overview of features that could potentially influence the linguistic credibility of a CPA and a starting point for further research of the variety.

Since a comprehensive account of Chinese English spoken by people from all over the country (see figure 1.2 for an overview of the linguistic diversity found in China) is not feasible, this thesis focuses on providing a stratified, quantitative description of two previously researched provinces (Guangxi and Yunnan) and adds research on a province that has not been covered by previous work (Guangdong). Adding data to the description of Chinese English spoken in Guangxi and Yunnan furthermore adds to the scientific rigor of the descriptions, as previous work was only carried out by one research group for each province.

The work presented on durational variation and the question of how varieties differ across Kachru's circles (cf. chapter 4) adds the first descriptions of durational variation in Czech English and Italian English. While there is slightly more previous work on Czech English and Slavic varieties in general, there is little previous work on Italian English. This thesis thus contributes to the quantitative description of these two varieties. Additional work on the vowel spaces of Czech and Italian English has been presented at the IAWE25 conference (Albrecht & Schmied, 2023), but is not included in this thesis, which focuses on Chinese English.

In summary, this thesis both tackles broader questions, such as the effect of a linguistically credible, non-native conversational pedagogical agent on learning outcomes in *Hybrid Societies*, in which humans and EDTs interact, and investigates language details, such as the vowel spaces and speech rhythm of Chinese English.

1.6 Methodological Considerations

1.6.1 Data Collection

Since the research presented here has been conducted as part of an interdisciplinary project, the data collection process had to be designed in a way to satisfy the requirements of the disciplines involved, in particular the requirements of neural TTS systems and sociolinguistics. While training data for neural TTS systems is usually collected by having participants read out extensive reading scripts, usually taken from public domain books (Isogai, Mizuno, & Mano, 2005), sociolinguistic research uses only comparatively short reading passages (Deterding, 2006a) and focuses on interview questions aimed at eliciting free speech (Tagliamonte, 2006, 2012). The content of the speech elicited does, however, matter little in both disciplines, which are interested in the language quality rather than the content. Furthermore, both training a neural TTS system and analyzing linguistic variation require a sufficient

number of phonemes. The amount of data required for training a neural TTS system still is much larger than the the amount of data used in linguistic analyses. To put this into perspective, a sociolinguistic interview may take up to two hours in total, whereas the LJSpeech data set (Ito & Johnson, 2017), which is used in many TTS applications contains 24 hours of speech of a single speaker. To reconcile the differing requirements, two different approaches were adopted. First, the TTS system was modified to use transfer learning, i.e. it is first trained on a larger data set to learn general features of English and at the end trained on a smaller data set to learn specific language features, in particular pronunciation and intonation. Second, our colleagues from computer science proposed a text selection algorithm (Taubert, Sternkopf, Kahl, & Eibl, 2022) to generate reading passages for TTS data collection that are as short as possible while maintaining a sufficient phoneme coverage of the target (English) language. In the final version of the data collection strategy, three linguistic reading passages amounting to about five minutes of speech and a reading script specifically designed for our TTS system that generated about 35-45 minutes of speech were used. Afterward, an interview based on questions provided in Tagliamonte (2006) with additional questions relevant to the target audience (i.e. questions about student life, or living in Germany) was conducted for 45-60 minutes, depending on the enthusiasm of the participants. In total, about 90-120 minutes of audio were scheduled to be collected.

Another important aspect of collecting data internationally is legal compliance. For the research presented here, both speech data and demographic data were collected. While within Germany and the European Union, such data collection is covered by the General Data Protection Regulation (GDPR), which regulates the rights of data subjects, outside of the EU, such rules are not as homogeneous. In the planning stage of our data collection procedures, the P.R. China introduced the Personal Information Protection Law (PIPL), creating a new, dynamically evolving legal situation. In a joint publication with the legal experts of the CRC *Hybrid Societies*, we developed best practice guidelines for data collection for research in China (S. Meyer, Albrecht, et al., 2022). By following privacy by design principles, in only collecting strictly necessary data and by keeping recordings and demographic data separate, only linked by a non-personally identifiable number, we ensured legal compliance with both GDPR and PIPL.

1.6.2 Data Collection and COVID-19

The research project within which the research presented in this thesis has been carried out started in early 2020. According to the project plan, a data collection trip to our partners at Sun Yat-sen University in Guangzhou, China was planned for the early stages of the project. However, the global COVID-19 pandemic made travel impossible for several years. To mitigate this threat to the success of the project, a twofold strategy was implemented.

Firstly, to maintain the strong focus on Chinese English, which is also central to this thesis, different possibilities to collect data despite restrictions on international travel were explored. In a first attempt, collecting data via online conferencing solutions was tested, however, the audio quality of the integrated recording function was not up to the quality standards required for phonetic analysis. To compensate for this issue, recording equipment was sent to our partners in China, however, China customs never cleared the parcel, despite import duties being paid, and returned it after keeping it for a year. The issue was eventually resolved by our Chinese partners purchasing recording equipment to carry out the data collection according to the specifications of the project. Ultimately, the data collection was only completed in early 2023, roughly two years behind schedule.

Secondly, while the efforts to collect data in China were ongoing and after meeting people in person and traveling within Europe were possible again, the scope of the project was partially widened. On the one hand, Chinese students at TU Chemnitz were recorded, to allow a pilot study on features of Chinese English and to test the TTS system. Even though the data was not ideal, because these participants had been exposed to German and were learning the language, it was still vital to test our methods and to train the TTS system. On the other hand, additional non-native varieties were included in the data collection and analysis. Two Nigerian scholars who were guest researchers at the Department of English Linguistics were recorded and in turn, helped with recording participants from the Nigerian diaspora living in Chemnitz. Additionally, Italian and Czech English data has been collected at partner institutions in Brno, Czech Republic, and Modena, Italy. Some of the data has been used in the speech rhythm analysis presented in chapter 5 and a detailed analysis of the different vowel spaces based on the automated analysis method described in section 1.6.4 has been presented at the IAWE25 conference in Stony Brook, New York (Albrecht & Schmied, 2023).

In conclusion, while the COVID-19 pandemic posed a severe risk to the project, the mitigation strategies implemented were successful and produced more than enough data for linguistic analysis. However, without the help of trusted, longstanding partners the data collection would not have been possible.

1.6.3 Data Transcription

In general, data can be transcribed with different levels of detail, depending on the research goal. The commonly used levels of granularity are orthographic, phonemic, and phonetic transcription (Schultze-Berndt, 2006). In an orthographic transcription, spoken language is usually represented in the standard orthography of the target language, sometimes with colloquialisms. In a phonemic transcription, basic phonemic contrasts (i.e. minimal pairs) are annotated. In contrast, a phonetic transcription typically contains all words in their phonetic form, usually in the International Phonetic Alphabet or SAMPA (Wells, 1997), a system devised to allow typing phonetic symbols with standard ASCII characters. It might also contain suprasegmental information on word stress and tone. Transcription is a time-consuming task, the industry standard for orthographic transcription of audio data seems to be at around a 1:4 ratio, meaning transcribing one hour of a recording takes about four hours. For a more detailed, phonetic transcription, this ratio can increase up to 1:200 (Brinckmann, 2017).

These ratios suggest that manually transcribing just the reading passage parts of our Chinese English data, which amounts to around 20 hours of audio for 30 participants, would require up to 4,000 hours for a phonetic transcription. Even the orthographic transcription of the interviews, roughly 30 hours of data would take around 120 hours. Additionally, there is the Czech (12 participants), Italian (10 participants), and Nigerian (7 participants) data, which in total amounts to roughly 58 hours of recordings. In total that amounts to 60 hours of reading data and 60 hours of interview data, corresponding to 480-24,000 hours of work in transcriptions. This amount of work would not be feasible for larger projects, let alone a small team of a single linguist and two to three student assistants.

In sociolinguistic research and phonetics, the transcription and annotation of data have traditionally been done manually (for general conventions of transcribing sociolinguistic data see Tagliamonte (2006), for details of data annotation and transcription in Praat see Brinckmann (2013)). However, with the rise of machine learning technology, new methods became available to researchers. Forced alignment tools such as MAUS (Schiel, 1999), FAVE (Rosenfelder et al., 2015), or the Montreal Forced Aligner (MFA

McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017b) use machine learning to automatically align audio and text, usually accepting plain text and audio files as input and providing time-stamped Praat TextGrid files as output. These tools generally rely on pre-trained models of a specific variety that should be aligned. unfortunately, most available models are for native varieties, except for a model for Nigerian English that is available for the MFA.

Consequently, a semi-automatic transcription pipeline was developed in close cooperation with the project partners from computer science, consisting of a set of scripts wrapping the MFA and providing several convenience functions for working with Praat TextGrid files (Taubert, 2023a, 2023b). The reading data and the interview data required slightly different approaches in transcription. While the data collected via the reading scripts could directly be aligned by the MFA, the interview data had to be orthographically transcribed before forced alignment. Due to the aforementioned time-consuming nature of manual transcriptions, only five-minute samples from the interviews of our Chinese participants were transcribed.

Before the actual transcription process could start, a pronunciation dictionary used by the MFA and our tools had to be generated. For the vast majority of words in the reading scripts and interviews, the International Phonetic Alphabet (IPA) transcriptions could be looked up automatically in dictionaries. Transcriptions of out-of-dictionary words were generated using a sequence-to-sequence model for grapheme-to-phoneme conversion (Taubert, 2022a). These automatically generated transcriptions had to be manually checked for correctness. Afterward, the actual four-step transcription process could start:

1. In the first step, the Montreal Forced Aligner was used to generate Praat TextGrid files from the plain text, orthographic transcriptions, using the previously generated and verified pronunciation dictionaries. The output of the MFA was also manually checked for several features of the recordings that would impede the rest of the automatic analysis: (1) stutter, for example at the beginning of words, as in *i-i-important*; (2) pauses in the middle of words, as in *impor...tant*; and (3) laughter, coughing, and other forms of non-speech, which were marked to be ignored in the analysis.

2. In the second step, the timestamps in the TextGrid files were verified. While the MFA produces good alignments for clear speech, especially the special features mentioned in step 1 often lead to misplaced word and phoneme boundaries.

3. In the third step, an additional annotation tier was added and pauses, stutters, and intervals in which the interviewer spoke were marked.

4. In the fourth step, the automatically generated IPA transcriptions were crosschecked with the audio, for example, if a normally reduced vowel was realized as a full vowel, or if a consonant in a consonant cluster was dropped.

The resulting files, as illustrated in figure 1.4, then formed the basis for both the linguistic analysis and the training of the neural TTS system. The figure shows the spectrogram generated from a recording of a participant and the accompanying TextGrid file containing the transcription. There are four layers in the TextGrid file: (1) the first layer contains time-stamped words; (2) the second layer contains automatically generated phonemes; (3) the third layer contains the manually annotated and aligned

spoken phones (note that the boundaries have been adjusted too); and (4) the fourth layer contains special markers, for example, to exclude utterances by the interviewer or inaudible utterances.

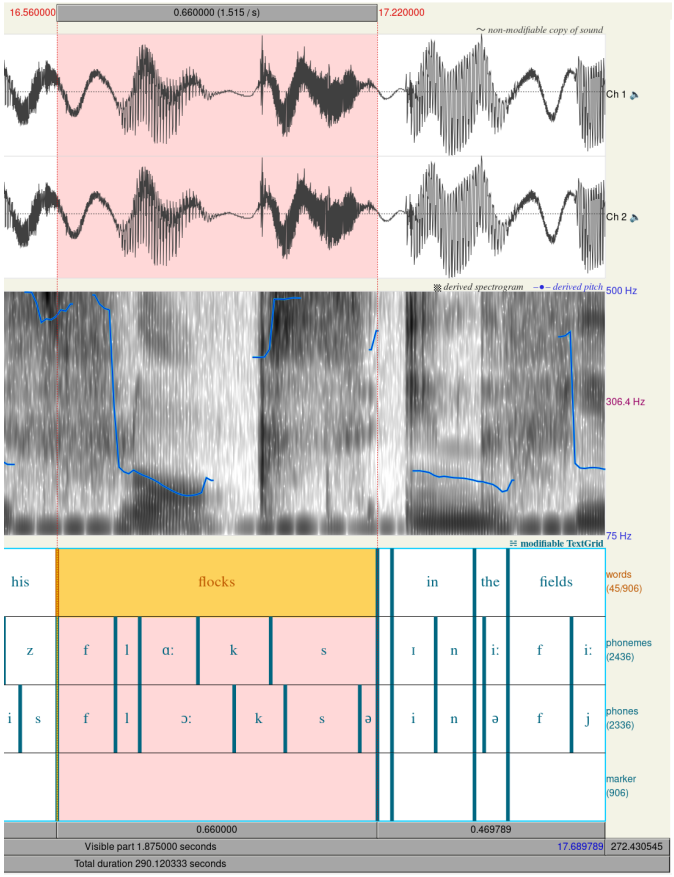


Figure 1.4: Example of a complete transcription in Praat according to the four-step process used in the project.

The procedure described in this section produces very accurately transcribed data while trying to minimize the workload for human transcribers by using automatic tools and scripts whenever possible. In addition, the transcriptions resulting from this procedure satisfy both the requirements of a thorough linguistic analysis and the training of a neural TTS system that allows fine-grained control of the linguistic features of the speech it produces. However, due to the challenges outlined in section 1.6.2, this procedure has only been applied to data collected before 2023. The data collected in 2023, meaning all 30 interviews of the participants from Guangdong, Guangxi, and Yunnan, were transcribed differently.

Initially, it was planned to force align the data of the 30 Chinese English speakers using the MFA and then adjust the linguistic analysis to work with the new data format, as by default, the MFA produces phonemes in the ARPABET (Rice, 1976), rather than the IPA. However, several issues were discovered with the recordings during the annotation process.

1. The instructions given to our colleagues had been interpreted curiously. As a general guideline

for the recordings, I had given rough estimates of how long the reading passages and interviews should last. However, the estimate that reading the reading passages for the TTS system should take about 35 minutes per participant had been interpreted such that every participant should read for 35 minutes and then stop. This resulted in 35-minute audio files with vastly different numbers of sentences read.

2. The reading skills of the participants were lower than expected. The majority of them had very low reading fluency, resulting in many false starts, lengthening of syllables to read ahead and sometimes even skipping difficult words.

3. Reading verbs proved especially challenging, as oftentimes the infinite form was read rather than a past tense form (e.g. *gave* was read as *give*).

While the MFA performs very well at word and phoneme levels, comparable to trained human transcribers, but with more alignment errors of above 25ms (MacKenzie & Turton, 2020; McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017a), the data suffering from these issues was beyond its comparabilities. Consequently, all vowels analyzed in the paper presented in chapter 4 were manually annotated.

1.6.4 Automated Analysis

The analysis of vowels in sociophonetic research can be done manually in Praat. However, this is only feasible for relatively low numbers of tokens per vowel per speaker. The data analyzed in the papers presented here typically contain several hundred to several thousand vowel tokens per speaker, with the largest data set, the American English reference data set used, containing 229,100 vowel tokens. A manual analysis is not feasible with data sets of this size.

Fortunately, Praat comes with scripting capabilities, that allow scripts to control every aspect of the program's functionality. Unfortunately, the documentation is lacking and existing scripts are hard to come by and often are handed down through the generations of researchers at linguistics departments. A side effect of this is that these existing scripts usually require data to be in a very specific format, which was a requirement at the time the original author wrote the script. However, the data format required by existing scripts that were available was not compatible with the data format described in section 1.6.3. With a bit of reverse engineering, it was possible to write a new script (Albrecht, 2023c) that was compatible with the data format used and that would automatically measure all the phonemes in the data, The script saves the phoneme measurement in a comma-separated values (CSV) file that can then conveniently be ingested in R.

The remainder of the analysis was carried out in R (R Core Team, 2023), using the tidyverse (Wickham et al., 2019), dplyr (Wickham, François, Henry, & Müller, 2020), phonR (McCloy, 2016), rstatix (Kassambara, 2023b), and e1071 (D. Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2022) packages. Plots were generated using the ggplot2 (Wickham, 2016) and ggpubr (Kassambara, 2023a) packages. The steps of the analysis differ slightly between the investigation of vowel spaces presented in chapter 3 and the investigation of speech rhythm presented in chapter 4. The general workflow follows the same pattern in both cases: load the data generated by the Praat script, transform the data and optionally filter outliers, calculate metrics, and plot results.

The vowel space analysis was produced according to the following procedure:

1. In the first step, the data generated by the Praat script was loaded into R, the data was cleaned by removing invalid measurements, a unique speaker ID was derived from file names, and the R data frames were transformed to make the data suitable for the other steps of the analysis process.

2. In the second step, the means of the first and second vowel formant frequencies were calculated, since the Praat script measures all phonemes at 20%, 50%, and 80% duration. Additionally, the formant frequencies were filtered for outliers using a Hampel filter (Hampel, 1974), which uses a movable window of plus/minus three absolute standard deviations. At this point, the filtered formant means were transposed to the Bark scale (for the research described in chapter 3 Traunmüller (1990)) or Lobanov normalized (for the research described in chapter 4 Lobanov (1971)).[13]

3. In the third step, the vowels were split into tense and lax subsets. For every tense and lax subset, the individual vowel formant frequency means and standard deviations of formant frequencies were calculated. Additionally, convex polygon hulls were computed from the vowel means for the tense and lax vowel data.

4. In the fourth step, the vowel space plots were generated, including scatter plots of all vowel measurements, the tense and lax vowel area polygons, individual vowel means, and error ellipses of one standard deviation.

5. In the fifth step[14], the influence of the sociolinguistic variables gender, province, and elicitation method on vowel realization was investigated. After testing the data for normality, using the Shapiro-Wilk test (Shapiro & Wilk, 1965), and for homoscedasticity (i.e. homogeneity of variance), using Levene's test (Levene, 1960), the Kruskal-Wallis test (Kruskal & Wallis, 1952) was selected as a non-parametric alternative to the ANOVA test, as it does not require normally distributed data and homogeneous variance.

6. In the sixth step, possible mergers were investigated using support vector machines in a binary classification task (similar to the approach described in H. L. Chan & Fuchs, fc.). The data sets were randomly split into a training data set containing 70% of the data and a validation set containing the remaining 30% of the data. The classifier was trained on the first two formant frequencies and the vowel label. Results were averaged across 50 iterations.

7. In the seventh step, possible mergers were investigated by comparing vowel areas of individual vowels and computing their overlap, using a modified version of the approach described in Albrecht et al. (2022).

For the research on speech rhythm presented, fewer steps were necessary:

1. In the first step, the data was transformed to contain only intervals either labeled as vocalic or consonantal, with adjacent consonantal intervals being joined and measured automatically

[13] For a detailed discussion of vowel formant normalization methods and the reasoning for choosing different methods for the two papers section 1.6.5.

[14] Note that this step requires a small amount of manual intervention. The results of the Shapiro-Wilk test and Levene's test have to be interpreted by the researcher. This might also be done before running the analysis. The type of data collected usually allows the researcher to anticipate the results of the tests.

in Praat. Then the data generated by the Praat script was loaded into R and transformed in preparation for the next steps. The data was split into data reading and interview data sets.

2. In the second step, the following speech rhythm metrics were calculated: ΔV, the standard deviation of vocalic intervals, ΔC, the standard deviation of consonantal intervals, and %V, the percentage of total speech duration that comprises vowel intervals (Ramus, Dupoux, & Mehler, 2003; Ramus, Nespor, & Mehler, 1999), the speech rate normalized versions of ΔV and ΔC, VarcoV and VarcoC (Dellwo, 2006), a pairwise variability index for vowels based on the durational differences between successive vowel intervals (i.e. successive syllables) nPVI-V, and rPVI-C (Grabe & Low, 2002; Low, Grabe, & Nolan, 2000). For a detailed discussion of these metrics see section 4.1.1.1.

3. In the third step, plots of combinations of metrics commonly found in the literature were generated.

Once the data annotation scheme and the scripts have been defined and tested, adding new data sets of different varieties for example, as was done in response to the challenges discussed in section 1.6.2, is a quick process. The only caveat is that the recordings have to be of sufficient quality to allow for successful forced alignment.

1.6.5 Vowel Formant Normalization

When a speaker produces a vowel, three different types of information are transmitted (Ladefoged & Broadbent, 1957): (1) phonemic information on the identity of the vowel that was produced; (2) physiological information on the size and shape of the speaker's vowel tract; and (3) sociolinguistic information. While all three types of information have been shown to influence formant frequencies (Labov, 2001; Peterson & Barney, 1952), physiological and sociolinguistic information have been considered unwanted variation in vowel perception studies (Peterson & Barney, 1952; Pols, Tromp, & Plomp, 1973), whereas physiological information has also been considered unwanted variation in sociolinguistic research (Adank, Smits, & van Hout, 2004). As a result, sociolinguists have employed several vowel formant normalization methods to eliminate this unwanted variation.

In general, there are four different goals of vowel formant normalization techniques that are difficult to achieve by a single algorithm:

1. to eliminate variation caused by physiological differences among speakers;

2. to preserve sociolinguistic/dialectal/cross-linguistic differences in vowel quality;

3. to preserve phonological distinctions among vowels;

4. to model the cognitive processes that allow human listeners to normalize vowels uttered by different speakers.

Thomas and Kendall (2007), quoting Disner (1980) and Thomas (2002), cited in Fabricius, Watt, and Johnson (2009)

For our overarching research goals outlined above, goals one and two are most important. The variation due to physiological differences, meaning differently sized vocal tracts due to height, age or sex, should be minimized. At the same time, sociolinguistically relevant features, i.e. the defining features of Chinese English or any other variety under investigation, should be retained.

Since the 1960s, several different vowel normalization procedures have been proposed. These procedures can be classified according to the types of acoustic information they utilize The most common classification is into vowel-intrinsic and vowel-extrinsic procedures (Ainsworth, 1975; Nearey, 1989). Vowel-intrinsic procedures, such as Bark[15] (Traunmüller, 1990), Mel[15] (O'Shaughnessy, 1987) and Syrdal and Gopal (1986) only utilize the acoustic information of a single vowel, wheres vowel-extrinsic procedures, such as Gerstman (1968), Lobanov (1971), Nearey (1978) and Watt and Fabricius (2002), utilize acoustic information of all vowels of a speaker. Furthermore, vowel normalization procedures can be classified as formant-intrinsic, i.e. utilizing only information of single formant, or formant-extrinsic, i.e. utilizing information of multiple formants (Fabricius et al., 2009). Lastly, normalization procedures can be categorized as speaker-intrinsic, i.e. utilizing information of one speaker, or speaker-extrinsic, i.e. utilizing information of several speakers (Adank et al. 2004, cited in Thomas and Kendall 2007). Due to the large variety of normalization procedures available, several studies (Adank et al., 2004; Clopper, 2009; Disner, 1980; Fabricius et al., 2009; Flynn & Foulkes, 2011) have investigated which are most suitable for sociolinguistic/sociophonetic research. The consensus of these studies is that vowel-extrinsic, formant-intrinsic perform best with regards to goals one and two. In particular, Lobanov performed consistently well according to all metrics.

Some researchers have cautioned that normalization might introduce artificial variation patterns not present in the raw data (Disner, 1980; Hindle, 1978), or that sociolinguistic variation might be lost (cf. goal two, Hindle, 1978). However, other studies showed that this is either not the case (Labov, 2001), or that the severity depends on the normalization procedure selected (Clopper, 2009; Fabricius et al., 2009). This debate is also linked to sample size, as normalization provides more benefits when comparing groups or speakers with relatively few tokens per vowel. Once a sufficient sample size is reached, even raw Hertz values can be comparable (Adank, van Heuven, & van Hout, 1999). Transforming raw Hertz values to the Bark scale, as in the study presented in chapter 3, might be a good compromise, whereas Lobanov normalizing might be preferable for small samples.

1.6.6 Vowel Length Normalization

While there is a variety of normalization methods that are relatively widely used to normalize vowel formants, the situation is different for vowel length. The rationale for normalizing vowel durations is similar to that of normalizing formant frequencies, that is to remove unwanted variation, either due to physiological differences between speakers or due to different speech rates. It is especially important to consider the influence of speech rate when sampling from connected speech (Wade, 2017). Several approaches to normalize vowel length have been proposed and used in previous research:

- segmental duration z-score normalization (Wassink, 2006),
- phrase-duration normalization (Wassink, 2006),

[15] Technically, Bark and Mel are nonlinear transformations of the frequency scale (in Hz), however, they are usually used in the same way as "real" normalization procedures.

- phoneme-word duration normalization (Paillereau & Chládková, 2019),
- combined method of speech rate normalization and Lobanov z-scores (Wade, 2017),

As with vowel formant normalization, vowel length normalization approaches have similar goals and thus the different approaches have been created for different purposes. Generally, approaches that normalize speech rate are appropriate when comparing groups of speakers, also between varieties. A combined approach with Lobanov (1971) z-scores might be useful when extending the analysis of the two-dimensional, normalized vowel space into the third dimension, as in Wassink (2006). While Lobanov normalized vowel lengths seem to be appropriate when comparing groups within and across varieties, they are slightly unintuitive, as they are no longer reported in Hertz (Hz), but rather as values on a Lobanov scale usually ranging from −3 to +3. Negative values on the scale indicate rather short vowels, whereas positive values indicate rather long vowels. This, however, might provide easy-to-read illustrations when investigating the tense-lax contrast.

In conclusion, there seems to be no best practice approach to vowel length normalization. While trying to eliminate the effect of speech rate through normalization seems to be generally desirable, z-score normalization should only be applied if the research design warrants it. As with vowel formant normalization, the argument that a sufficiently large sample size eliminates the need for normalization can be made.

1.6.7 Text-to-Speech System

The neural text-to-speech system developed in the larger research project that the research presented in this thesis was conducted in is based on two state-of-the-art components: Tacotron (Shen et al., 2018; Y. Wang et al., 2017), a neural model that predicts mel spectrograms from sequences of input symbols, and WaveGlow (Prenger, Valle, & Catanzaro, 2019), a neural vocoder that synthesizes these mel spectrograms into audio signals.

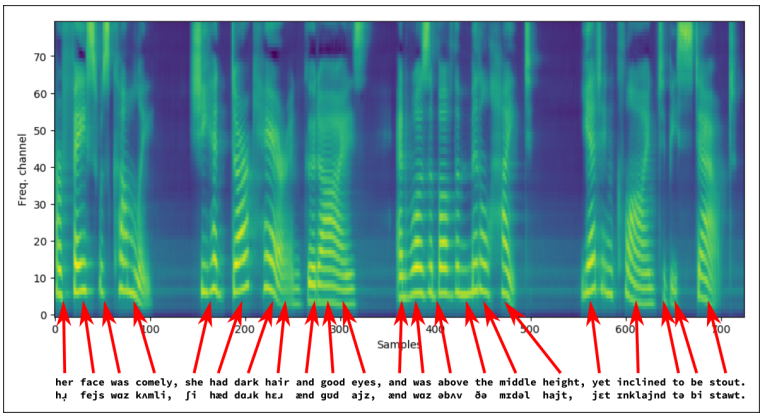


Figure 1.5: Example of a mel spectrogram with speech labels. Image courtesy of Stefan Taubert.

The Tacotron model builds on recent advances in machine learning and image recognition and learns to associate graphical representations of audio, in the form of mel spectrograms, with labels, representing words and letters in standard English orthography, as illustrated in figure 1.5. For the research

project *Credibility Through Non-native Language Varieties in Conversational Pedagogical Agents* (cf. section 1.5), the model has been modified to learn the relation between mel spectrograms and language represented by IPA symbols. To further improve the model, we extended the architecture, which already contained speaker embeddings, to also support separate learning of stress using a 1-hot vector for a better synthesis of the phonemes. For the vector, we used the value `0` as padding, `1` for consonants and punctuation, `2` for unstressed vowels, `3` for vowels with primary stress and `4` for vowels with secondary stress. The vector was concatenated to the symbol embedding before passing it to the encoder.

Tacotron was trained for 500 epochs with a batch size of 64 and a learning rate of 0.001. The training took about 100 hours and was done on Ubuntu 20.04 LTS using an Nvidia Titan RTX graphics card (with CUDA 11.6) and an AMD Ryzen 7 3800X processor. For WaveGlow version three of a pre-trained model provided by Nvidia (NVIDIA, 2019) was used, which was trained on the LJ Speech dataset (Ito & Johnson, 2017).

For the non-native varieties, the LJ Speech model was used as a base pre-trained model, and training weights of phonemes were mapped to their closest equivalent in the non-native model. Then, training was resumed for another 5000 epochs while reducing the learning rate exponentially by factor 0.97 from 0.001 to 0.00001 for fine-tuning the models. The same batch size of 64 was used. In addition, for the Chinese English model, phoneme duration was learned explicitly.

1.7 Synopses

Current research on the linguistic features of Chinese English

This paper argues that Chinese English is an expanding circle variety in the nativization stage and that there are central features of Chinese English regardless of a speaker's Chinese L1 or dialect.

After a brief overview of competing terminology used to describe the variety of English spoken by people from China, *Chinglish*, *China English*, and *Chinese English*, the term *Chinese English* is adopted, following Kirkpatrick and Xu (2002); Wei and Fei (2003); Xu (2006) and defined as follows:

> *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.

Using Chinese English instead of the previously popular term China English also reflects the discussion in the literature (Xu, 2017; Xu, Deterding, & He, 2017), that the negative association of Chinese English with Chinglish is disappearing and in the long run, the term Chinese English is more appropriate.

In the first part of the review, research on the phonology of Chinese English is summarized. The key findings from the literature include (1) accounts of individual vowel variation (Ao & Low, 2012, 2016: e.g. STRUT is pronounced as LOT in Yunnan, variation of the PRICE and FACE vowels); (2) epenthetic vowels, which usually means an extra schwa is added after a final plosive and before the next word, a common feature in Central and Northern China (Deterding, 2006b; Ho, 2003); (3) the widespread absence of vowel reduction (Ao & Low, 2012; Deterding, 2016); (4) vowel insertion in consonant clusters (Ao & Low, 2012); and (5) substantial variation in the realization of consonants (Ao & Low, 2012; Deterding, 2006b, 2016; S. Li & Sewell, 2012). Building on these previous studies, an in-depth analysis of Chinese English vowels of speakers from Guangdong, Guangxi, and Yunnan is presented in chapter 3.

In the next section, previous work on prosodic features of Chinese English is reviewed. The focus here is on the reported syllable-timed rhythm of Chinese English (Deterding, 2006b) in contrast to the traditional stress-timed rhythm of inner circle varieties of English. On a related note, it has also been shown that Chinese speakers tend to stress phrase final pronouns (S. Li & Sewell, 2012) and to stress the final syllable in multisyllabic words (Bian, 2013). The question of whether Chinese English is stress-timed or syllable-timed is addressed in chapter 4.

In the following section, features of Chinese English lexis are summarized, which is the aspect of the variety that has the longest tradition in research, with early accounts ranging back to the 1970s (Knowlton, 1970). Broadly speaking, Chinese English lexical items fall into one of three categories: (1) Chinese loanwords in English (Xu, 2010); (2) Chinese nativized English words (Xu, 2010); and (3) common English words used in Chinese English (Xu, 2020) Lexical items in category one usually either originate from Mandarin, such as *fengshui* or *guanxi*, or from Cantonese, such as *box choy* or *dim sum*. This category also includes loan translations, such as *Cultural Revolution* or *Red Guard*. Items in category two include English words that have changed in denotation and connotation, such as *propaganda* or *cadre*, which have a rather positive connotation in Chinese English, whereas *individualism* has a rather negative connotation (Xu, 2020).

In the next section of the review, syntactical features of Chinese English are discussed. The variation here does not always conform to (native) English norms without being ungrammatical (Z. Du, 1998). Instances of this are a preference for complex nominalization or premodification (Y. Liu, Fang, & Wei, 2017). Xu (2008) classifies syntactic variation into three categories: (1) preference, which refers to features that exist in the native model variety but are used with a different frequency or in different situations; (2) innovation, which refers to features that are generalized from marked usage in the model variety to unmarked use in Chinese English; and (3) transference, which refers to the rendering of L1 syntactic rules in English. Lastly, article usage has also been studied, mostly from a second language acquisition background (e.g. Chuang, 2005; Leroux & Kendall, 2018; Snape, Leung, & Ting, 2006; L. White, 2008), assuming some form of L1 transfer (for a detailed review and discussion see Albrecht, 2017).

In the last section of the review, discourse and pragmatic features of Chinese English are summarized. These features can be seen as a manifestation of Chinese cultural and social norms and include the "ancestral hometown discourse" (Xu, 2020) and other topics, but also variation in discourse patterns, such as the deviation from the general-particular pattern (Y. Jiang, 2002).

Following the comprehensive review of previous work on Chinese English, the features are synthesized into a suggested pronunciation model and a categorization of phonological features following the eWAVE categorization model. Lastly, shared features with the related varieties of Hong Kong English and Singapore English are briefly discussed. The article closes by highlighting the need for large-scale studies on the features of Chinese English to give the variety a status similar to other Asian varieties of English.

The Chinese English Vowel Space Across Three Provinces

In this paper, a quantitative analysis of the Chinese English vowel space is presented, researching the influence of the variables gender, province, and data elicitation method (reading passages vs. sociolinguistic interview) on the vowel space of English speakers from Guangdong, Guangxi, and Yunnan. Chinese English vowels have been shown to vary in their vowel space and their lax-tense distinction (cf.

chapter 2). However, studies on regional varieties of Chinese English are rare, especially with stratified participant samples.

Therefore, data from 30 participants, 15 male and 15 female, from the provinces of Guangdong, Guangxi, and Yunnan has been collected at Sun Yat-sen University in Guangdong, China. Participants were asked to read linguistic reading passages (The North Wind and the Sun, The Boy Who Cried Wolf, and Arthur the Rat, for a discussion of these reading passages see Deterding, 2006a), a list of sentences optimized for training neural text-to-speech systems (for details on how this list was created see Taubert et al., 2022) and interviewed using a question catalog based on Tagliamonte (2006). While the initial plan was to use the automated transcription approach described in section 1.6.3, it turned out to not be feasible due to the participants' low reading and speaking fluency, which resulted in high amounts of false start, hesitation, and even reading different words than in the script. Ultimately, the transcription had to be done manually. For the reading passages, ten tokens of each vowel have been transcribed per participant, and for the interview, all vowel tokens in five-minute samples have been annotated.

The analysis has been carried out in three stages: (1) vowel area plots of tense and lax vowel vowels have been plotted in Bark for the male and female groups from each province and aggregated data sets; (2) the influence of the variables elicitation method (reading/interview), gender, and province have been tested using the Kruskal-Wallis test, as a non-parametric alternative to the ANOVA test, as the data did not satisfy the conditions of normality and homoscedasticity; and (3) potential mergers (FLEECE-KIT, DRESS-TRAP, and GOOSE-FOT) were investigated using both an approach based on the metric described in chapter 5 and an approach using a machine learning approach (support vector machines in a binary classification task).

The results show that the data elicitation method has an almost negligible influence on the vowel spaces of the participants. Similarly, the variable province only had a small influence on the vowel spaces of the participants. The variable gender, however, had the most significant influence on the vowel spaces of the participants. The tense and lax vowel areas for both individual groups and aggregated data sets have a relatively high size overlap with the tense vowel spaces (about 56%). Furthermore, evidence suggests mergers of the FLEECE-KIT and GOOSE-FOOT vowel pairs for the majority of our participants. The findings are consistent with previous studies and demonstrate that the Chinese English vowel space shows clear and distinctive non-native characteristics.

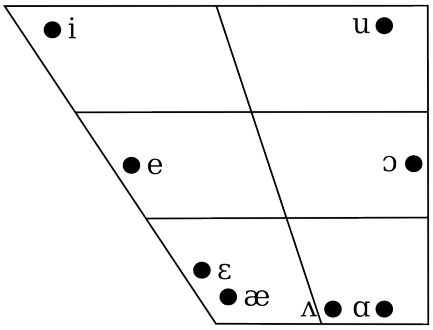


Figure 1.6: Proposed Chinese English vowel inventory

In conclusion, it has been shown that the non-native vowel spaces of English speakers from the Chinese

provinces Guangdong, Guangxi, and Yunnan have a consistent shape, with a very weak tense-lax contrast and likely mergers of the FLEECE-KIT and GOOSE-FOOT vowel pairs. Of the three sociolinguistic variables, only gender had a significant effect on the vowel spaces of the participants with a large effect size. Based on these findings, a tentative vowel inventory, as presented in figure 1.6 has been proposed.

Durational Variation across inner/outer/expanding Circle Varieties of English

In this paper, a quantitative analysis of durational variation in English varieties across the three variation circles (Kachru, 1985) is presented. We investigate four inner circle varieties (American, British, Scottish, Irish English), one outer circle variety (Nigerian English), and three expanding circle varieties (Czech, Italian, Chinese English).

Starting with the earliest accounts (Pike, 1945), languages have conventionally been categorized as either stress-timed, with feet of equal duration or syllable-timed, with syllables of equal duration (Abercrombie, 1967: 97). The third category of mora-timed languages (cf. Ladefoged & Johnson, 2014) has been abandoned in recent research (Fuchs, 2023a). However, as research on speech rhythm progressed, it became clear that the binary categories could not be supported empirically (Dauer, 1983), giving rise to the notion of speech rhythm as a gradable phenomenon (L. White, Mattys, & Wiget, 2012) and to the reconceptualization of speech rhythm as durational variation (L. White & Malisz, 2020).

While the definition of rhythm in language is already challenging (see chapter 4 for different proposed definitions), there is empirical evidence from language production that the underlying assumption of isochrony of language elements (either syllables or feet) is not tenable (Bertinetto, 1989; Roach, 1982). A possible explanation of the perceived isochrony that lead to the idea of speech rhythm categories might be found in structural phenomena, such as vowel reduction or syllable structure (Dauer, 1983). The combination of high onset and coda complexity, reduced vowels in weak syllables, and many strong syllables due to primary and secondary stress in English create a steep "prominence gradiant" in English (Nolan & Asu, 2009) that is perceived as apparent stress-timed rhythm (L. White & Malisz, 2020). Consequently, the observable durational variation is only a surface phenomenon of underlying langauge structure.

In variationist research, inner circle varieties of English have traditionally been categorized as stress-timed, whereas outer and expanding circle varieties have been categorized as syllable-timed (see contributions to Kortmann & Schneider, 2004), despite the conceptual shift from speech rhythm towards research on durational variation.

Based on the concept of durational variation, we investigate the influence of language experience, learner background, region, and data elicitation method on the durational variability observed in outer and expanding circle varieties of English, using inner-circle varieties as points of reference. Starting from the impressionistic description of these varieties, we use quantitative durational variation metrics (DVMs) to measure durational variation to see whether the descriptions are mirrored in the values produced by the metrics.

While the inner circle varieties analyzed here have consistently been categorized as stress-timed, the situation is not as clear for the other varieties. The outer circle variety under investigation, Nigerian English, has been studied extensively, in terms of phonology (Gut, 2008), prosody (Gut, 2005), and morphology and syntax (Alo & Mesthrie, 2008). It has to be kept in mind, however, that Nigeria has a very diverse glssl1 landscape, making the existence of subnational varieties likely, especially for the languages with the most speakers, Hausa, Igbo, and Yoruba (Gut, 2005: 38). A previous quantitative

study of Nigerian English speech rhythm placed the variety more towards the syllable-timed end of the continuum, more so than British English, but less than other outer circle varieties. As for the expanding circle varieties, Chinese English has traditionally been classified as syllable-timed (cf. chapter 2), which has been quantitatively confirmed in recent years (Ding & Xu, 2016; A. Li & Post, 2014; S. Liu & Takeda, 2021). There is hardly any research on Italian English in general and no research on Italian English speech rhythm in particular. Assuming L1 transfer, looking at research on Italian might provide a first intuition, which has been classified as less stress-based (than English for example, see Bertinetto, 1981; Ramus, 2002). The situation with Czech English is similar, with little previous research being available. While Volín and Skarnitzl (2010) did not explicitly investigate whether Czech English is stress-timed or syllable-timed, they reported that Czech English was perceived as more strongly foreign accented in their less stress-timed condition. Research on L1 Czech is unfortunately also inconclusive, as the language has initially been classified as syllable-timed (Duběda, 2004), but as stress-timed more recently (Dankovičová & Dellwo, 2007).

The analysis has been carried out according to the procedures described in section 1.6.3 and section 1.6.4. The following durational variation metrics (DVMs) were calculated: ΔV, the standard deviation of vocalic intervals, ΔC, the standard deviation of consonantal intervals, and %V, the percentage of total speech duration that comprises vowel intervals (Ramus et al., 2003, 1999), the speech rate normalized versions of ΔV and ΔC, VarcoV and VarcoC (Dellwo, 2006), a pairwise variability index for vowels based on the durational differences between successive vowel intervals (i.e. successive syllables) nPVI-V, and rPVI-C (Grabe & Low, 2002; Low et al., 2000).

Our findings show that metric results differ between inner, outer, and expanding circle varieties. The results of the expanding circle varieties are more diverse than the results of the inner circle varieties. Interestingly, Czech and Italian English produced different results in most metrics, strongly suggesting L1 influence. Both the raw values and the figures plotting conventional combinations of vocalic and consonantal measures suggest that varieties could fall into broad groups. Previous work on DVMs indicated possible influences of articulation rate and data elicitation method on metric results. While we observed differences in values, these differences were not found to be statistically significant, even though articulation rate in particular differed visibly between reading and interview data.

Comparing our results with previous work, the majority of the values generated by the DVMs are in similar or plausible ranges. Our nPVI-V values for the Chinese English data are slightly higher than values reported in the literature, however, we suppose this might be the influence of language proficiency and fluency, which was low in our data. The automated analysis pipeline utilizing speech rhythm metrics has proven to be a useful tool in quantifying impressionistic classifications of varieties as either stress-timed or syllable-timed.

Towards a Vowel Formant Based Quality Metric for Text-to-Speech Systems

To evaluate the speech produced by (neural) text-to-speech systems, several objective and perceptual metrics have been proposed, which mostly focus on intelligibility and naturalness. In this context, intelligibility means the clarity of the speech, whereas naturalness refers to the similarity of the speech to a human voice and the general ease of listening, global stylistic consistency, regional or linguistic nuances (Sotelo et al., 2017; Taylor, 2009a).

Some of the objective metrics available include the mel cepstral distance (MCD) (Kubichek, 1993) and F0-RMSE or voice/unvoiced mismatch (VUV) (Kuo, Ouyang, Aryal, & Lanchantin, 2019), while

subjective metrics, meaning metrics using human evaluators, include mean opinion scores (MOS) (Chu & Peng, 2006), preference tests with side-by-side comparisons (Chevelu, Lolive, Maguer, & Guennec, 2015), intelligibility rate (IR) (Ping et al., 2018) and MUSHRA (*Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, 2015).

While using objective metrics to evaluate synthesized speech is preferable as these metrics are more time and resource-efficient, subjective metrics still outperform them and Consequently are most widely used (Chevelu et al., 2015). While the gold standard of subjective metrics in evaluating synthesized speech is not going to change anytime soon, the approach described here can further reduce evaluation costs by providing a novel objective metric capable of estimating the naturalness of monophthongs. Additionally, the metric provides detailed insights into the quality of synthesized audio and provides pointers for further optimization of the system.

The metric proposed here combines the automatic transcription process using the Montreal Forced Aligner, described in section 1.6.3, and the automatic generation of vowel area plots and overlap percentages, described in section 1.6.4. It furthermore applies the approach to different subsets of the data, namely the vowel space areas spanning the means, the central 75% of data points, and all data points. Additionally, the metric is computed for the whole vowel space, as well as for all individual vowels. The metric is designed to evaluate synthesized audio for which the ground truth, i.e. an original recording is available.

The results of a pilot study show that both the vowel space calculated from phoneme averages and the vowel space calculated from the central 75% of data points overlap greatly at 93.20% and 97.70% respectively. The overlap percentage of the vowel spaces covering all data points is lower at 91.50%. The overlap percentages of individual vowels differ substantially, with percentages ranging from 56.54% to 98.22%. As stated above, the metric reveals areas in which the synthesis is not yet optimal and could be improved by additional training of the system.

The main limitation of the metric is the reliance on the accuracy of the forced alignment software, which, however, performs at a level comparable to human transcribers (MacKenzie & Turton, 2020; McAuliffe et al., 2017a). Given a sufficient number of data points, which usually is the case in TTS applications as a substantial amount of training data is required, and applying some sort of filter for outliers, this limitation is offset.

In conclusion, the vowel formant-based quality metric for TTS systems proposed here can serve as a basis for developing more comprehensive and linguistically sound quality metrics for TTS applications. Extending the metric to include diphthongs and vowel stress and combining it with research on the perception of synthesized speech, correlating the values generated by the metric with scores by human evaluators might further reduce the need for participant-based perceptual evaluations of TTS systems.

Competent but Cold: Learning with a Pedagogical Agent Speaking Non-Native Varieties

Most embodied digital technologies used in educational contexts, as pedagogical agents for example, and more broadly most text-to-speech products, only support inner circle varieties of English (Karakaş, 2017). With the increasing digitization of education and increased use of English as a medium of instruction, non-native speakers of English now have more access to educational resources than ever. Yet, the language they hear in educational resources, especially those available online, is often different from the variety they have been exposed to in their home country, i.e. their own variety. While adapting to unfamiliar varieties is of course possible (Cristia et al., 2012), it requires exposure time, which otherwise

could be used for subject matter learning, ultimately putting non-native learners at a disadvantage in the global competition.

Building on previous research that found instruction in students' variety of L1 German, improved learning outcomes (Rey & Steib, 2013), this study investigates whether a similar effect can be found in L2 English spoken by a pedagogical agent[16].

For human teachers, the influence of non-native English on learning outcomes is correlated with attitudes towards non-native speech (Ahn & Moore, 2011). While there is no comparable research on synthesized speech, it has been shown that users prefer robots speaking their familiar variety (e.g. Andrist, Ziadee, Boukaram, Mutlu, & Sakr, 2015; Sandygulova & O'Hare, 2015; Tamagawa, Watson, Kuo, MacDonald, & Broadbent, 2011). Pedagogical (and conversational) agents are perceived as social partners (Louwerse, Graesser, McNamara, & Lu, 2009) and as such, are attributed characteristics and personality traits based on social and linguistic cues they send (Nass & Lee, 2001a). Consequently, agents should be constructed to evoke favorable social perceptions. If voice is the only mode of interaction an agent has, it should be optimized accordingly (Sutton et al., 2019). As evidence of the effectiveness of pedagogical agents in improving learning outcomes is mixed (Martha & Santoso, 2019), this study contributes to the growing body of research by investigating both the learning outcomes and the attitudes of users towards the pedagogical agent.

To investigate the efficacy of a pedagogical agent at improving learning outcomes, results from an online learning task administered to students at partner universities in Italy and the Czech Republic were analyzed. In total, the data of 74 Italian and 15 Czech students, who were mostly young (76% were 21-23 years old), female (70 females, 19 males), and studying in the field of Humanities (Czechs 80%, Italians 85%). Most of them were in a Master's program, except one Bachelor's and nine PhD students in the Italian sample. On average, they had studied English for a comparable amount of time (Czechs 16.9 years, SD = 4.53; Italians 15.2 years, SD = 2.81). The survey furthermore collected background information on participants' exposure to English and their technology affinity.

The learning material presented to participants was adapted from a Linguistics article by Delibegović Džanić and Berberović (2021). The text explains cognitive content mapping based on the joke "If life gives you lemons, a simple plastic surgery can give you melons" (Delibegović Džanić & Berberović, 2021: 124). The language used in the text was designed to be challenging, given the complex structure and academic vocabulary due to the academic nature of the topic. The text was synthesized using the text-to-speech system developed as part of a wider research project (cf. section 1.5) and had a length of 2:30 minutes, which is appropriate in terms of attention span and memory capacity, given that the material is only presented as audio (S. D. Craig & Schroeder, 2017). Participants were randomly assigned to listen to the lecture in either American English or their local variety. Afterward, they were asked five academic vocabulary questions and three transfer questions.

The results show that the variety spoken by the agent does not significantly improve learning outcomes. Regarding participants' attitudes towards the agent, the voice was consistently rated higher on COMPETENCE characteristics than WARMTH characteristics. In contrast to the Czech participants, the Italian participants consistently rated the agent speaking American English higher than the agent speaking Italian English

Ultimately, the study did not confirm the hypothesized improved learning outcomes due to a

[16] It is important to stress that the English language is only used as a medium of instruction. This study and the whole project (cf. section 1.5) are not concerned with language learning.

pedagogical agent speaking the students' variety of English. With an improved TTS system and a larger sample size, a future study might produce different results. Nevertheless, the finding that participants consistently rated the agent as competent but cold can already inform designers of inclusive digital learning technologies that adapt to the learners' individual needs.

2 Current research on the linguistic features of Chinese English

This chapter has been published as: Albrecht, S. (2021). Current research on the linguistic features of Chinese English. *World Englishes*, 42, 487-506. https://doi.org/10.1111/weng.12572.

Abstract

This paper argues that there are central features of Chinese English regardless of a speaker's Chinese first language (L1) or dialect. The current state of research on Chinese English is reviewed, outlining phonological, lexical, syntactical, prosodic, and discourse and pragmatic features of Chinese English. These features are categorized according to their pervasiveness based on different L1 backgrounds, to show that there are core features of Chinese English. The argument is limited by the low number of quantitative studies and the absence of large-scale quantitative studies.

2.1 Introduction

Over the last decades, English in China or Chinese English has seen considerable coverage in academic literature. In the early days, research focused mainly on English in China, meaning English in second language acquisition and educational contexts (see Bolton (2003) for a detailed historical account). Building on the works of Labov (summarized in Labov (1994, 2001, 2010)) and the theoretical works on variation by Halliday, McIntosh, and Strevens (1964) and Weinreich et al. (1968) and on world Englishes and leaving behind the colonial past of the English language (e.g. Bolton & Kachru, 2006; Kachru, 1996; Kirkpatrick, 2007; McArthur, 2002; Phillipson, 1992, 2010, 2018), the focus of research shifted from English in China to Chinese English, that is treating the English used by people with Chinese as their L1 in their own right.

Considering this notion, one of the core issues of researching Chinese English becomes apparent: the term Chinese English presupposes that there is a coherent language "Chinese" that is the shared L1 of all speakers of Chinese English. However, this is far from the truth. In fact, China has a very diverse linguistic landscape consisting of different dialects that are unique in their variety and complexity (Ramsey, 1989: 6). Most of the dialects belong to the Sino-Tibetan language family with some exceptions mostly found in Northern Central China and Northwest China. From a Western perspective, "Chinese", as in the language spoken in China, usually denotes Putonghua ("common language"), the official national language in the PR China, which is closely modeled on Beijing Mandarin (Ramsey, 1989: 8). A second dialect that people in the West might know about is Cantonese (*guangdonghua*), originating from Guangzhou (also known as Canton), which might commonly, but incorrectly, be used to refer to the entire Yue language family (Norman, 1988: 242). Most importantly, the majority of these varieties are not mutually intelligible (Norman, 1988: 213).. In particular, the considerable phonological variation between different dialects introduces more variables when studying Chinese English spoken in areas with a complex linguistic landscape, such as Guangxi and Yunnan, assuming one's preferred theory of second language acquisition allows for L1 interference (Full Access Full Transfer Hypothesis (Schwartz

& Sprouse, 1994, 1996; Sprouse & Schwartz, 1998), Full Access Hypothesis (Epstein et al., 1996, 1998), Failed Functional Features Hypothesis (Hawkins & Chan, 1997)). Coincidentally, the two provinces mentioned above have been subject to the most scrutiny in previous investigations of Chinese English. The linguistic landscape in China is undergoing changes as well, mostly due to government campaigns promoting the use of Putonghua, leading to Yang (2009) arguing that other Chinese varieties are in decline. Furthermore, due to the role of Chinese English as a norm-dependent (Kachru, 1998), Expanding Circle variety, codification and nativization processes (see for example E. W. Schneider (2014b)) are ongoing. Thus, differentiating between an accepted feature and a developmental mistake is a matter of frequency and attitude and is therefore difficult, especially since many of the studies that will be reviewed only recruited relatively small numbers of participants from small communities (such as a single university). There is another dimension to the fact that many studies recruited their participants from a single university. While good universities (high ranking) attract students from different provinces, the different social strata of a society are not equally represented at these higher education institutions. In addition, it has to be kept in mind that English still only plays a minor role in most Chinese people's daily lives (Bolton & Graddol, 2012).

Finally, it has to be noted that in addition to research published in English, there have been numerous publications in Chinese since the 1980s[1] (e.g. R. Du & Jiang, 2001; Feng, 2013; Ge, 1980; G. Gu & Xiang, 1997; Hu, 2011; J. Huang, 1988; W. Li, 1993; L. Sun, 1989; T. Sun, 2006; R. Wang, 1991; Xie, 1995). For an overview of these works, see Xu (2017, 2020). Unfortunately, there are no full translations available, and sometimes not even the Chinese original text can be sourced. For a comprehensive review, see Xu (2017).[2]

2.1.1 Nativization

Kachru (1992a) established four functional uses of a language that are crucial in judging the status of a variety: (1) instrumental, (2) regulative, (3) interpersonal, and (4) imaginative/innovative. The instrumental function is most prominently realized in the education system, with many universities in China now using English as a medium of instruction and offering courses in English (Ma & Xu, 2017: 192). The regulative function, meaning the use of Chinese English in administration and law, is less frequently employed in comparison to the other functions, but efforts are being made by the government in setting up English government websites and by releasing official English translations of legal texts (Ma & Xu, 2017: 194). The interpersonal function is realized by speakers of Chinese English in communicating with migrants in China and on the Internet (Ma & Xu, 2017: 191-192v). The imaginative and innovative function of Chinese English can be observed in literary texts (c.f. H. Zhang, 2002) and also in the classroom (Ma & Xu, 2017: 192). Butler (1997) identified five criteria to establish whether a new variety is legitimate: (1) a distinctive pronunciation handed down from one generation

[1] While I have taken great care to include all studies on Chinese English, it might be possible that I have missed some, particularly if published in Chinese only.

[2] Xu identifies four phases of research by Chinese scientists on Chinese English: "the "enlightenment" (1980–1997), the "great leap forward" (1998–2001), the "renaissance" (2002–2012), and the "open-door" (2013 onwards)" (Xu, 2017: 260). Furthermore, Xu lists a number of major research areas: "whether Chinese English exists; naming Chinese English; defining Chinese English; distinguishing between Chinese English and *Chinglish*; the *China English* debate; awareness of and attitudes towards Chinese English; acceptability of Chinese English; functions and significance of Chinese English; features of Chinese English; applications of Chinese English; implications of Chinese English for ELT in China; reviews of Chinese English research; stages of researching Chinese English, and methods for researching Chinese English" (Xu, 2017: 260)

to another, (2) a unique lexicon, (3) a history of using the variety, (4) literature written in that variety, and (5) reference tools (Butler, 1997: 106 cited in Ma and Xu 2017). Ma and Xu (2017: 195) argue that all five criteria are satisfied to varying degrees and therefore Chinese English can be situated at the nativization stage of the dynamic model by E. W. Schneider and Kirkpatrick (2010).

2.1.2 Definitions

When discussing the topic of English spoken by Chinese native speakers in China, there are a few terms that can cause confusion as they are often times used interchangeably or with varying definitions: *Chinglish*, *China English* and *Chinese English*. Out of the three, *Chinglish* has probably the most widely agreed upon definition, as an interlanguage between Chinese and English that is usually manifested in the form of literal translations from Chinese. In other words, *Chinglish* can be defined as Chinese syntax with English words and strong L1 interference in pronunciation, resulting in low levels of intelligibility (Wei & Fei, 2003: 43, cited in Eaves 2011). See example 1 for an illustration. Also note that the term *Chinglish* carries a negative connotation, but is sometimes used for humorous purposes by speakers of Chinese English.

(1) Good good study, day day up!
hao hao xuexi, tian tian xiangshang
'Study hard: do have a progressive spirit every day'
(adapted from Wei & Fei, 2003)

Chinese English was first been coined by Ge (1983), who considered it the outcome of translating items specific to Chinese culture into English (Wei & Fei, 2003: 43). Eaves (2011: 66) defined *Chinese English* as "the interlanguage spoken by Chinese learners of English". However according to Wei and Fei (2003), this term has attracted a similar amount of criticism as *Chinglish*, as it also carries a negative stigma. Since both *Chinglish* and *Chinese English* are perceived to have a negative connotation, W. Li (1993) defines *China English*, a term introduced by Ge (1980)[3], as a term "for an English with Chinese characteristics and culture" (Wei & Fei, 2003: 44). Kirkpatrick and Xu (2002) provide a full quote of this definition (the original paper was published in Chinese):

> China English is based on a standard English, expresses Chinese culture, has Chinese characteristics in lexis, sentence structure and discourse but does not show any L1 interference. (Kirkpatrick & Xu, 2002: 269, cited in Eaves 2011)

Out of these competing terms for English spoken by people in China, I will use the term *Chinese English* to describe the emerging variety. Xu (2008) tentatively defines Chinese English as follows:

> [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 at varying levels of language, and it is used primarily by Chinese for intra- and international communication. (Xu, 2006: 287, cited in Xu 2008)

Since both definitions presented above contain parts that I consider important, I define *Chinese English* as follows, merging the definitions by Wei and Fei (2003, cited in Kirkpatrick and Xu 2002) and Xu (2006).

[3] I would like to thank one of the anonymous reviewers for pointing out that it was in fact Ge (1980) who first introduced the term *China English*.

> *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.

Using Chinese English instead of the previously popular term *China English* also reflects the discussion by Xu et al. (2017) and Xu (2017), who argue that the negative association of Chinese English with *Chinglish* is disappearing and in the long-run the term Chinese English is more appropriate.

2.2 Phonology

The phonology of Chinese English has so far been the main focus of research, based on the number of articles written on the topic. I have identified 24 features of Chinese English that are relevant within our constraints outlined above.

2.2.1 Vowels

Due to the size and linguistic diversity of China, only a limited number of regions and provinces have been studied. In particular, there are studies on the larger regions of Northern and Central China (Deterding, 2006a; Ho, 2003) and on the provinces of Guangxi (Deterding, 2016, 2017) and Yunnan (Ao & Low, 2012, 2016). The most salient feature of Yunnan English, appearing in 90 per cent of the population, is the realization of the (British English) STRUT vowel (ʌ) as an (American English) LOT vowel (ɑ) (Ao & Low, 2012: 31). This leads to the word *duck* being pronounced as /dɑk/. Another feature that is well-documented in the literature is epenthetic vowels, which usually means an extra schwa is added after a final plosive and before the next word (Deterding, 2006b: 180).. This feature is widespread in Central and Northern China (Deterding, 2006b; Ho, 2003) and in Yunnan (Ao & Low, 2012), however less frequently found in Guangxi (Deterding, 2016). The latter finding is interesting, given that this feature seems to be common in different Asian varieties of English such as Korean English (e.g. Darcy & Thomas, 2019) or Japanese English (e.g. Nomura & Ishikawa, 2018; Yazawa, Konishi, & Kondo, 2015) and that S. Li and Sewell (2012) reported it for speakers coming from both the north and south of China. There are two other types of epenthetic vowels that received slightly less attention in the literature (Ao & Low, 2012), namely insertion of a vowel into a consonant cluster within a word and the insertion of a schwa before a nasal /n/ within a word. These different types of variation provide compelling evidence that epenthetic vowels are a feature of Chinese English.

The absence of reduced vowels, especially in function words, is widespread in Northern and Central China, Yunnan (Ao & Low, 2012: 31-32), and Guangxi (Deterding, 2016: 15). The feature is especially well-pronounced in function words, where reduced vowels are realized as full vowels in 89 per cent of the cases (Deterding, 2016). Particularly in Guangxi, vowels are sometimes nasalized, sometimes even if there is no nasal consonant (Deterding, 2006b: 184). However, this feature seems to be dependent on the mother tongues of the speakers and requires further investigation (Deterding, 2016). Another interesting feature that is predominantly found in Guangxi is overlap in the length of the FLEECE and KIT vowels (Deterding, 2016: 14). Diphthongs are another area in which variation can be found in both Yunnan and Guangxi. In Yunnan, the FACE vowel /eɪ/ is often realized as [ɑɪ], [iː] or [e], leading to *raising* being pronounced as ['rɑɪzɪn] and *safety* as ['siːftɪ] or ['seftɪ]. The PRICE vowel /ɑɪ/ is often realized as [e], meaning *while* can be pronounced [wel] or *tried* can be pronounced [tred] (Ao & Low,

2012: 32). Both /eɪ/ and /ɑɪ/ vary in about 50 per cent of the occurrences (Ao & Low, 2016). It has to be noted that variation of the FACE vowel /eɪ/ also occurs in native varieties of English (Wells, 1982: 141-142). Furthermore, the variation of the PRICE vowel /aɪ/ can be classified as monophthongization, a common feature of non-native varieties that has been reported, for example, for African Englishes (Schmied, 1991a).

Ao and Low (2016) observed two other notable forms of variation that occur much less frequently. They report that the NEAR vowel /ɪə/ is realized as a SQUARE vowel [ɛə] in about 20 per cent of the occurrences in their data and /ɛə/ is sometimes realized as a NURSE vowel [ɜ:]. S. Li and Sewell (2012) report diphthong shortening (or monophthongization) for speakers coming from Northern China (S. Li & Sewell, 2012: 87-88). Notably, the shortening only occurs if the diphthong is followed by a consonant. This could mean that the shortening is due to the influence of L1 phonotactics, as Mandarin has similar diphthongs that cannot, however, be followed by a consonant (S. Li & Sewell, 2012: 88). Standard Chinese syllables are structured CGVX, with the final X is a glide or consonant (Duanmu, 2007: 71). However, the possible consonants are further restricted to /n/, /ŋ/ and /ɻ/ (Chao, 1968: 51) Concluding the survey of different forms of diphthong variation, it appears that variation of diphthongs is a feature of Chinese English. However, the specific form that this variation takes might be dependent on variables such as L1 or L1 dialect and phonetic environment.

2.2.2 Consonants

Chinese English consonants are more interesting than vowels, at least as far as the amount of variation that has been reported in the literature is concerned. Most of the variation fits into either of the following two categories: substitution (a phoneme is realized as a different phone) or omission (the phoneme is not pronounced at all).

One of the most frequently found forms of consonant variations in many non-native and even in some native varieties of English is variation of the voiceless dental fricative /θ/. In Chinese English, the voiceless dental fricative is mostly realized as a voiceless alveolar fricative [s]. This variation is widespread in Northern and Central China, where it occurs in 71 per cent of the observed cases (Deterding, 2006b: 194), and in Yunnan, where it appears in roughly 50 per cent of the observed cases, but attribution to sociolinguistic variables is difficult (Ao & Low, 2012: 31). In Guangxi however, this type of variation rarely occurs (Deterding, 2016: 6). S. Li and Sewell (2012) observed similar variation of the voiceless dental fricative /θ/ being realized as [s], with a tendency for appearing in word final position and an overall low frequency of occurrence. Similarly, the voiced dental fricative /ð/ also varies in Chinese English and is mostly realized as a voiced alveolar fricative [z] or as a voiced alveolar plosive [d] (Deterding, 2006b: 188). The variation seems to differ by region with both [d] and [z] being found in Northern and Central China (Deterding, 2006b: 188), whereas in Yunnan [z] seems to be the dominant variation, even though ethnic background seems to have an influence (Ao & Low, 2012: 31). In Guangxi, /ð/ is realized as [z] in 79 per cent of the observed occurrences (Deterding, 2016: 7). S. Li and Sewell (2012) observed a 55 per cent ration of /ð/ being realized as [d], similar to Hung (2005, cited in Deterding 2006b). A possible explanation of this phenomenon could be the influence of Hong Kong English, since S. Li and Sewell collected their data from mainland Chinese students studying in Hong Kong, and the variation of /ð/ as [d] has been reported by Deterding, Wong, and Kirkpatrick (2008) for Hong Kong English. Another interesting variation that can also be found in native varieties is the realization of the voiceless glottal fricative /h/ as a voiceless velar fricative [x]. This kind of variation might be

attributed to the fact that the [x] sound is represented as 'h' in the Pinyin spelling of Chinese words (W.-S. Lee & Zee, 2003) and is mostly found in Northern and Central China (Deterding, 2006b: 190). A possible explanation of this variation might be found in the fact that in many Chinese classrooms writing is traditionally favored over speaking (Wen, 2018: 527), even though this is changing (X. Zhang & Ardasheva, 2019; Zheng & Cheng, 2008). A feature that seems to be exclusive to Chinese English spoken in Yunnan is the realization of the voiceless postalveolar fricative /ʃ/ as a voiceless palatal fricative /ç/ (Ao & Low, 2012: 31). The feature has a frequency of about 40 per cent and can be observed in words containing the digraph 'sh,' like *shepherd*.

The voiced postalveolar fricative /ʒ/ also varies as [ɹ] (Deterding, 2006b: 191) or [j] (Ao & Low, 2012: 31). The variation can be observed particularly frequently in words such as *usual*, *usually* or *pleasure*. There seems to be a certain regional diversity to the variation, with the realization as[ɹ] mostly found in Northern and Central China as Deterding (2006b) reports in accord with Hung (2005). This report is furthermore confirmed by S. Li and Sewell (2012), however they argue that the [ɹ] sound has a retroflex quality, probably due to an influence of Northern Chinese dialects, and might be more accurately transcribed as [ɻ]. For Yunnan, Ao and Low (2016) reported a frequency of 60 per cent of the occurrences of /ʒ/ varying as [j], which has not been reported in other studies. In Guangxi, this type of variation does not seem to be very prominent (Deterding, 2016: 5). Continuing with fricatives, both the voiced labiodental fricative /v/ and voiced alveolar sibilant /z/ vary in Chinese English. The voiced labiodental fricative in words such as *traveler* is often realized as the voiced labio-velar approximant [w]. This is one of the most salient features of English spoken in Guangxi (Deterding, 2016: 5), however it is not very common in Yunnan (Ao & Low, 2012: 32). E. W. Schneider (2011: 151) confirms the low frequency of this variation. The voiced alveolar lateral approximant /l/ is involved in three different types of variation: (1) it can be realized as [n] (Deterding, 2006b: 192), (2) it is part of the l-r continuum (Deterding, 2016: 10), and (3) it can be vocalized (Deterding, 2006b: 192). The first variation of /l/ is the realization as a voiced alveolar nasal [n], which is rare but salient in Northern and Central China (Deterding, 2006b: 192). The second variation, different realizations of sounds on the l-r continuum, mostly in consonant clusters, seems to be regarded as a common variation in the general perception of Chinese English; however, the data shows that it only appears rarely in Guangxi (Deterding, 2016: 10) and it is not reported for other parts of China. Vocalized /l/ is only possible for occurrences of dark ɬ in a syllable coda. It seems to be the most common phenomenon of the three and has been reported to be widespread in Northern and Central China (Deterding, 2016: 11) and in Yunnan (Ao & Low, 2012, 2016), where up to 90 per cent of the occurrences varied. The vocalization can take different forms, for example, *wolf* can be pronounced [wɒ], [wɔːf] or [ʊf] and *full* can be realized as [fɔː], [fuː] or [fʊə] (Ao & Low, 2012: 31 contains more examples). Lastly, the dark /l/ sound [ɬ] that is often omitted in English is spoken in Guangxi (Deterding, 2016: 11). To discern which of these /l/ variations are features of Chinese English and which are carried over from the model language is difficult, since Wells (1982: 258-259) already reported /l/ vocalization in British English.

Omission of single word final consonants is a common variation reported by Deterding (2016: 10) for speakers from Guangxi, exemplified by the omission of the word final /t/ in *hot afternoon*. A possible variant of this phenomenon is the replacement of a single word final consonant with a glottal stop, as in *duck*, where /dʌk/ becomes [dʌʔ] (Chang, 1987 cited in Deterding 2016, Chang 2001; Deterding 2016: 10). However, it is likely that these variation patterns are influenced by whether the consonant in question is word final or clause final. Therefore, the fact that Deterding (2016) did not confirm the findings of Chang (1987, 2001) regarding the word *duck* might be because it appears only in clause final

position in the wolf passage used by Deterding (2016). A special case of this variation is the omission of word final /n/ that is often combined with the insertion of a schwa, both a widespread feature of English spoken in Yunnan (Ao & Low, 2012: 31). The omission of word final /n/ leads to words such as *soon* being realized as [suː] and the schwa insertion results in *afternoon* being realized as [ɑːftəˈnuːən] (Ao & Low, 2012: 31). Omission of word final consonants is not limited to single consonants, but also occurs in consonant clusters. Simplification of word final consonant clusters is a widespread feature of English varieties, not just non-native varieties (Schreier, 2005: 27 cited in Deterding 2016). Deterding (2016: 8) reported an omission rate of 32.9 per cent in word final /st/ consonant clusters for students from Guangxi. Furthermore, the variation seems to be more likely to occur if the next word starts with a vowel. A more specific case of final consonant cluster simplification was observed by S. Li and Sewell (2012). They report that in 55.2 per cent the final /t/ or /d/ in past tense forms are omitted. They speculate that voiceless plosive + plosive clusters (for example /pt/ in /dʒʌmpt/ or /zd/ in /siːzd/) are especially challenging for Chinese learners and warn that they also observed substantial inter-speaker variation (S. Li & Sewell, 2012: 93). Thus, instead of an omission of a word final consonant, these cases could also be classified as consonant cluster simplification. E. W. Schneider (2011: 149) identified a strong aspiration of voiceless stops, in particular /t/ in his data of mainly Cantonese speaking mainland Chinese students at the Chinese University of Hong Kong. The aspiration can be observed in initial, medial, and word final position and even in consonant clusters. This feature appears in combination with vowel epenthesis and consonant cluster reduction.

Lastly, rhoticity is a topic of debate in the study of Chinese English, mostly because it is unclear what the model is. Some scholars claim that American English serves as the model (E. W. Schneider, 2014b: 19), while others point out that British English is the (official) model (C.-C. Cheng, 1992; Y. Jiang, 2002). Data from research does not confirm the role of American English as a model (Deterding, 2016: 12) for Guangxi and Ao and Low (2016) adopt British English as a model for their study of English in Yunnan, citing C.-C. Cheng (1992) who states that British English has served as the model of English in the Chinese education system since the 1950s. S. Li and Sewell (2012) observed a blend of American and British pronunciation features in their data, even within a single word, for example *tomato* pronounced with British [t] sounds but with an American [ɐɪ]. Rhoticity might also vary regionally, either due to historical developments, or due to L1 phonological preferences, such as the use of [ɹ] by speakers of northern Mandarin dialects. An important caveat that has already been hinted at above is the fact that for some of the variations presented here, it is not clear whether the variation is a feature of Chinese English or due to ongoing linguistic changes in the model variety. Possible cases of this include variation of the FACE diphthong and /l/ vocalization, which have been reported by Wells (1982).

2.3 Prosody

One of the most frequently listed prosodic features of Chinese English is a syllable-timed rhythm in contrast to the typical stress-timed rhythm of English (Deterding, 2006b: 183). This feature becomes especially apparent in multisyllabic words (He & Li, 2009: 72). Furthermore, Chinese English has a tendency to stress final pronouns, like textithim in the phrase *around him* (Deterding (2006b: 194), S. Li and Sewell (2012: 93)). This feature seems to be widespread in northern and central China (Deterding, 2006b: 194) but not in Guangxi (Deterding, 2016, 2017). S. Li and Sewell (2012) confirm this feature, even though they say it is slightly less prominent. Regarding word stress, Deterding (2016: 16) observed that in multisyllabic words, *concern* in particular, the stress is on a different syllable. However, even in

Guangxi this feature is not very widespread. Similarly, Bian (2013) observed that Chinese university students have a tendency to stress the last syllable in multisyllabic compounds. Regarding the differences in patterns of speech, a number of studies compared tone languages to stress languages. G. T. Chen (1972) reported wider f0 ranges and larger standard deviations for Mandarin speakers in contrast to English speakers. Eady (1982) compared English and Chinese and reported a greater amount of dynamic movement for Chinese. Xue, Hagstrom, and Hao (2002) observed lower minimum f0 and larger f0 ranges in younger Chinese-English bilinguals, but no such difference in older bilingual speakers. A detailed review of studies on crosslinguistic differences can be found in Dolson (1994).

Keating and Kuo (2012) compared speaking fundamental frequency in English and Mandarin and report slight differences for certain text types. However, it is not clear from their paper whether the fundamental frequency values were normalized before comparison, hence the results should be interpreted cautiously. In accord with previous studies, they observe a wider f0 range of Chinese speakers (Keating & Kuo, 2012: 1058). Ding, Hoffmann, and Hirst (2016) conducted a similar study, investigating normalized f0 patterns in English, Chinese, and Chinese English. They found that "only pitch range on the phoneme level and pitch change amount on the utterance level are consistently significantly different between English and Chinese, and between English and L2 English for both males and females" (Ding et al., 2016: 759). Chinese English speakers exhibited a higher pitch range on the phoneme level, which can be attributed to L1 transfer, as they might attach tones to syllables in English (Ding et al., 2016: 759). The pitch change amount was also higher for Chinese English speakers, which Ding et al. attribute to the fact that native English speakers reduced more unstressed vowels and spoke faster in general (Ding et al., 2016: 759).

Since there has been little research on prosodic features of Chinese English, it might be worth looking at some findings from analyses of Hong Kong English. Setter (2003) found that Hong Kong English speakers show less difference in the relative syllable duration of tonic, stressed, unstressed, and weakened syllables than British English speakers (Setter, 2003: 467). In a recent study, Ran, van de Weijer, and Sloos (2020) compared the intonation of Cantonese speakers of English from Hong Kong and Guangzhou. They find that the intonation patterns of Hong Kong English and Guangzhou English do not differ significantly, but deviate from native varieties of English. In particular, content words are consistently accented (Ran et al., 2020: 735), possibly due to their lexical importance. Furthermore Ran et al. report that there is no global rising in yes-no questions or echo questions and that the pitch accents of words are subject to down-step and boundary tones (Ran et al., 2020: 735). Other researchers have also shown that Hong Kong English has a distinct tonal quality (Cheung, 2009; Gussenhoven, 2012; Luke, 2000; Wee, 2008). It has to be kept in mind that Hong Kong English is largely influenced by Cantonese and thus it remains to be seen as to how far these observations apply to Chinese English.

2.4 Lexis

Early accounts of Chinese borrowings in English include Knowlton (1970) and Cannon (1988) (cited in Xu, 2020), which they define as "[w]hen Chinese speakers of English refer to things Chinese" (Kirkpatrick & Xu, 2002: 270-271 cited in Xu 2020). The lexis of Chinese English can be categorized into three groups: (1) Chinese loanwords in English, (2) Chinese nativized English words, and (3) common English words used in Chinese English (Xu, 2020: 269), of which the first two are of particular interest, especially since Xu (2020) does not give examples of words from category three. The three categories are notably a non-exhaustive subset of "refer[ing] to things Chinese". Chinese

loanwords in English (category one mentioned above) stem either from Cantonese, for example *bok choy*, *dimsum*, or from Mandarin (or Putonghua), for example *fengshui* and *guanxi*. Moreover, loanwords can also be loan translations, such as the *the Cultural Revolution* and *Red Guard* (Xu, 2020: 269). He and Li (2009: 73) note that Chinese loan translations in English are not standardized and a single Chinese phrase can be rendered in up to 16 different ways in English. Yang (2009) observes that early Chinese loanwords often stemmed from Cantonese, while newer loanwords mostly originate from Mandarin. This change mirrors the growing importance of Mandarin due to language policy and the rising influence of the Pinyin transliteration system. Newer loanwords also occur in a wider range of semantic fields (Yang, 2009: 105). In addition to these 'standing' loanwords, Xu (2020: 269-270) identifies ad hoc loanwords, which are the spontaneous usage of Chinese words in English if they have no adequate English equivalent. Chinese nativized English words (category two mentioned above) are also interesting, because their original meanings have changed semantically both in denotation and connotation through semantic broadening or narrowing and amelioration or pejoration (Xu, 2020: 270). Early accounts of this phenomenon include Chang (2001), who states that the word *propaganda* has a neutral or even positive connotation in Chinese English, whereas it is usually perceived as negative in Inner Circle varieties. Other examples of words with changed denotation include *cadre* and *migrant worker*, words with changed connotation including *individualism* and *comrade* (see Xu, 2020: 27-28 for a detailed discussion).

In contrast to previous research, which has mostly been qualitative, Liang and Li (2017) did a quantitative study on Chinese English collocations, in particular collocations of *all-around*. Using corpus linguistic methods, they show that all-around has different collocations in Chinese English than in British English and that the majority of Chinese English collocations are not found in British English (Liang & Li, 2017: 69). Their research shows that collocations of *all-around* have been nativized in Chinese English and are used differently than in British English (see example 2).

(2) The quality of the teaching staff should be improved in an all-around way.
(Liang & Li, 2017)

Liang (2015) identifies a number of collocations that are overused in Chinese English as well as a number of collocations relating to sociopolitical ideology. Many of these, especially of the latter type, can be regarded as loan translations from Chinese. In addition, Liang (2015) observes a semantic preference for words and collocations with positive connotations in Chinese English. Liang reports that for example *people* collocates more frequently with positive words, such as *increase*, *save*, *improve*, in Chinese English than in British English (Liang, 2015: 231). It would be interesting to further investigate the parts of speech of these collocations. In this context, exploring differences in idiomaticity, metaphor, and reference context might provide further insights. Studies on other varieties of English have shown that words are used in different reference contexts, for example specific terms are used in more general contexts (Schmied, 1991b: 88-89). Xu (2020) also notes that the distribution of Chinese English lexical items differs by text type or even style. In literary texts, more loan translations can be found and, also, literary texts written in Chinese English contain translations of Chinese idioms and proverbs (Xu, 2020: 271).

2.5 Syntax

The syntax of Chinese English has received little attention in research so far. An early account by Z. Du (1998) highlights that Chinese English sentence structure does not always conform to English norms without being ungrammatical, as illustrated in example 3, in which the first sentence could be seen as less idiomatic as the second.

(3) *Yesterday Mary bought a book.* vs *Mary bought a book yesterday.*
(Z. Du 1998: 11, cited in Xu 2017)

Y. Liu et al. (2017) analyzed nominalization in two corpora of media texts from China and Britain. They found higher frequencies of complex nominalization in the Chinese English media text, whereas the British English media texts contained more simple nominalizations (Y. Liu et al., 2017: 89). In addition, they showed that the Chinese English texts rely more on premodification, whereas the British English texts favor postmodification (Y. Liu et al., 2017: 89). The preference for premodification in Chinese English might be explained by the fact that Chinese usually uses left-branching structures with premodification, whereas English has more options for right-branching structures (Y. Liu et al., 2017: 90). When discussing pre- and postmodification, the existence of distinct word classes is a necessary precondition. However, for some languages, like Chinese, there are clear distinctions between some word classes, like nouns and verbs, whereas for other word classes, like verbs and adjectives, there is no consensus whether they are distinct or not. Earlier works tend to hold that verbs and adjectives are not separate word classes in Chinese (McCawley, 1992), whereas more recent works take the opposite stance (Paul, 2005). For a comprehensive crosslinguistic overview of this issue see Haspelmath (2012). Another explanation, in particular for the preference of simple nominalization in the British English texts, might be found in the ongoing process of colloquialization in British English, which has been widely reported in the literature (see for example Biber 2003; Hundt and Mair 1999; Leech, Hundt, Mair, and Smith 2009 cited in Y. Liu et al. 2017).

hen analyzing syntactic variation, Xu (2008) differentiates three different types of variation: (1) preference, (2) innovation, and (3) transference. Preference refers to features that exist in the native model variety (such as British or American English), but are used less frequently or only in certain situations. Innovation refers to features that are generalized from marked usage in the model variety to unmarked usage, for example unmarked use nominalization in Chinese English (Xu, 2008: 7) or unmarked use of *is it?* and *isn't it* it with rising intonation in Singaporean English (Weber & Platt, 1980: 80, cited in Xu 2008). Transference refers to the rendering of (syntactic) rules from the L1 in L2 production, such as the subclause preceding the main clause in Chinese English (Xu, 2008: 7).

Xu (2008) identifies numerous features of Chinese English syntax:

(1) spoken: deviant syntactic expressions
    a) Adjacent default tense (ADT)
    e.g. *Last year, I write a letter.*
    b) Null-subject/object utterances (NS/O)
    e.g. *Sometimes Ø just play basketball.*
    c) Co-occurrence of connective pairs (CCP)
    e.g. *Though we need money, but freedom is maybe the first priority.*

d) Subject pronoun copying (SCP)
e.g. *One of my roommates, he found it in ...*

e) Yes-No response (Y/NR)
e.g. *A: So, you don't like sports. B: Yeah.*

f) Topic comment (TC)
I think being a teacher, the life will be too easy.

g) Unmarked OSV (OSV)
e.g. *Probably some other kinds of jobs I also want to try.*

h) Inversion in subordinate finite wh-clauses (ISC)
e.g. *I really don't know what is International English.*

(2) written, formal

a) Nominalization
e.g. *The crowded trains, planes and buses attest to the high volume of people* ***embarking*** *on the sacred journey home.*

b) Multiple-coordinate construction
e.g. *To close the economic gap, top official agreed yesterday to [deepen the ongoing reforms on the grain distribution system], [further restructure the agriculture sector] and [regulate agricultural business, by making it more efficient and structured].*

c) Modifying-modified sequencing
e.g. *Although it is not a requirement, most participants are currently staying in China, either receiving an education or working.*

(3) written, informal

a) Use of imperatives
e.g. *She turned to the soldiers. 'Tell your leader we thank him.'*

b) Tag variation strategy
e.g. *'He just wanted to have a good time. That was all. It's not like adultery or bigamy, is it?' 'But it's a crime that got him put in jail,' I said.*

(adapted from Xu, 2008, 2010)

W. Jiang (2017) studied the modified-modifying sequence in compositions by Chinese advanced users of English. In particular, the subordinate conjunctions although, because, if, and when were investigated. The data shows that these conjunctions only appear in initial and final position, in line with the description of standardized English by Biber, Johansson, Leech, Conrad, and Finegan (1999), who stated that only a small number of adverbial clauses appear in medial position, contrasting Quirk, Greenbaum, Leech, and Svartvik (1985). W. Jiang (2017: 104) states that this is likely a feature of Chinese English. Furthermore, the data shows that "a majority (78.2%) of the sentences follow the order of given information preceding new information" (W. Jiang, 2017: 104). Lastly, it seems that in complex clauses in which the subordinate clause is linked by *although*, *if* or *when*, the sentence follows the modifying-modified sequence (the modifying clause in initial position), like Chinese. Whereas in

complex clauses in which the subordinate clause is linked by *because*, the sentence follows the modified-modifying sequence (the modifying clause in final position), like English. W. Jiang (2017: 104) attribute this to L1 influence, in line with Xu (2010).

He and Li (2009) summarize previous research on Chinese English syntax by Chinese scholars in four points.

1. Idioms made up of four morpho-syllables: many Chinese idioms consist of four syllables, which are often used as (literal) translations in Chinese English

   (4) 事半 功 倍
   shì bàn gōng bèi
   'effort halved, result doubled'

2. Parallel structure: often used in Chinese in expressing words of wisdom

   (5) 吃一堑，长 一智
   chī yī qiàn, zhǎng yī zhì
   'a fall into the pit, a gain in your wit'

3. Topicalization of adjuncts: modifiers, such as adverbials or adverbial clauses, are generally placed in front of the main verb of a sentence.
   e.g. *This morning I bought a book.*

4. The Null Subject parameter: in contrast to English, Chinese is a pro-drop language that does not require a subject in a sentence.
   e.g. *Very glad to write to you again.*

(adapted from He & Li, 2009: 73-74)

Article usage in Chinese English has received some attention from a more second language acquisition focused background (see for example Chuang, 2005; Leroux & Kendall, 2018; Snape et al., 2006; L. White, 2008). The consensus seems to be that article usage in Chinese English varies in comparison to standardized varieties of English, however the theoretical explanations of this fact differ. While some attribute the variation to different settings of the Article Choice Parameter (Ionin, 2003; Ionin, Ko, & Wexler, 2004), others investigate the influence of hearer knowledge and specific reference (Leroux & Kendall, 2018). They all have in common the assumption of L1 transfer from Chinese, which is traditionally considered a language without articles, though some scholars argue that there is an ongoing grammaticalization process in Chinese (P. Chen, 2003, 2004; C.-T. J. Huang, 1987; X. Li & Bisang, 2012). Since previous research treated article variation in Chinese English as a feature of learners' interlanguages, it remains to be seen if it can be confirmed using variationist sociolinguistic approaches, similar to Hong Kong English (Wong, 2020).

Lastly, the substitution of *he* and *she* by Chinese speakers of English is a rather curious phenomenon that can be linked to L1 influence as the corresponding words in Chinese (他 'he' and 她 'she') are both pronounced tā (Deterding, 2017). This feature occurs with a frequency of about 15% in the data analyzed by Deterding (2017).

2.6 Discourse and Pragmatic Features

Given the intercultural nature of Chinese English, it can be expected that, apart from purely linguistic transfer, there will also be discourse and pragmatic transfer. Xu (2020: 276-277) identifies a number of interesting features, notably in literary texts. The first feature is the so called "ancestral hometown discourse" (Xu, 2020: 276). Chinese speakers are likely to inquire about others' and share about their own hometown, as it constitutes part of their identity. The second feature is that "elements of Chinese discourse patterns are interwoven in the text in an almost seamless fashion" (H. Zhang 2002 cited in Xu 2020). Six types of Chinese discourse patterns can be identified: (1) discourse of "political status" and "political life", (2) discourse of "law" and "social order", (3) discourse of "power' and 'hierarchy", (4) discourse of "*guanxi*" and "backdoor practice", (5) discourse of "work unit" and "welfare", and (6) discourse of "face" and "name and honor" (Xu, 2020: 276-277).

Lastly, the third feature is a pragmatic feature, involving assumptions generally shared by speakers of Chinese and expectations built on cultural discourses (Xu, 2020: 277). This includes use of Chairman Mao's quotes, loan translations of swear words, and address terms. Native English varieties usually structure discourse in a general-particular pattern (*theme* and *rheme*, see Halliday & Matthiessen, 2013). Chinese English discourse rarely follows the general-particular pattern (Y. Jiang, 2002: 10, see also M. Wang and Li 1993). M. Wang and Li reported that, in their study, Chinese students used the problem-solution pattern most frequently, rather than the general-particular pattern (M. Wang and Li 1993 cited in He & Li, 2009: 74). A possible reason for this might be the transfer of L1 discourse structure, as (Kirkpatrick, 1993, 1996) showed that information in Modern Standard Chinese is usually sequenced in a justification-claim, or modifier-modified, sequence, particularly in because-therefore constructions, in contrast to the claim-justification sequence utilized in English.

Ren (2017) investigated pragmatic strategies of Chinese university students with different language levels in writing emails expressing gratitude and found that students consistently choose strategies that differ from strategies employed by speakers of inner circle varieties of English. In particular, Chinese students choose more formal greetings and closings, reflecting pragmatic strategies for status-unequal communication in Chinese (Ren, 2017: 121, see also Y. Gu 1990) and strict adherence to templates taught in language classes. Furthermore, Chinese students used "repayment" discourse strategies, which are considered important in American English (Eisenstein & Bodman, 1986) only infrequently, confirming research by S. W. Cheng (2005) (Ren, 2017: 121). Lastly, Ren (2017: 116) showed that higher proficiency level students write significantly longer emails expressing gratitude, which seems at odds with J. Liu and Ren (2016, cited in Ren 2017) who found that lower proficiency level students write longer apology emails.

2.7 Discussion

The sections above have shown that there has been a substantial amount of research on Chinese English. Further more, as time progressed, the research aims have shifted from defining Chinese English and establishing it as a genuine variety, to investigating its linguistic features. The features reported in the literature have to be approached with caution, given the diverse linguistic landscape of China and the fact that many studies are impressionistic and even the quantitative studies usually are very small-scale. Even though a comprehensive description of Chinese English is not yet possible, identifying trends and recurring features in these studies might still provide valuable insights into possible general features of

Keyword	Symbol	Confidence
KIT	i	high
DRESS	ɛ	medium
TRAP	ɛ	low
BATH	ɑ	medium
LOT	ɔ	low
CLOTH	ɔ	low
STRUT	ɑ	high
FOOT	u	low
FLEECE	i	high
GOOSE	u	medium
PALM	ɑ	medium
START	ɑ	medium
NURSE	ɜ	medium
NORTH	ɔ	medium
FORCE	ɔ	medium
THOUGHT	ɔ	medium
NEAR	ɪə	medium
SQUARE	ɛə	medium
CURE	ʊə	medium
FACE	ɑɪ	high*
PRIDE	e	high*
VOICE	ɔɪ	medium
MOUTH	aʊ	medium
GOAT	oʊ	low
HAPPY	i	medium
LETTER	ə	medium
RABBIT	i	medium
ADDED	i	medium
BEAUTIFUL	ʉ	low
PIANO	i	medium
AGO	ə	medium
BECAUSE	i	medium

Table 2.1: Pronunciation of Vowels in Chinese English Keywords (based on *Pronunciation Model: Hong Kong English*, n.d.)

Chinese English and also provide directions for future research. One of the most consistently reported features is the absence of reduced vowels, particularly in function words. A likely reason might be that vowel reduction does not exist in Chinese. The overlap of the FLEECE and KIT vowels in Chinese English might be a related phenomenon, since neither standard Chinese nor Cantonese have the KIT vowel, the only possible exception being Shanghai Chinese (Y. Chen & Gussenhoven, 2015). A reason for this could be the absence of vowel length contrasts in general, potentially affecting the GOOSE and FOOT vowels as well.[4]

In absence of a complete description of the Chinese English vowel system, it might be worthwhile

[4] I would like to thank one of the anonymous reviewers for pointing out this possibility.

to consider a related variety, Hong Kong English. Table 2.1 attempts a synthesis of the Chinese English vowel system based on previous research presented above and work on Hong Kong English found in the online edition of the Oxford English Dictionary (*Pronunciation Model: Hong Kong English*, n.d.). The confidence column indicates whether a vowel is a feature of Chinese English, Hong Kong English, or British English. The values are: (1) *high*, the available data strongly suggests that the vowel varies in Chinese English; (2) *high**, there is some data that suggests variation of this vowel but not strongly; (3) *medium*, this vowel is the same as in the British English model but there is no Chinese English data available; (4) *low*, this vowel varies in Hong Kong English and there is no Chinese English data available. Unfortunately, data is available for only four vowels, highlighting the urgent need for further work on the description of the Chinese English vowel system.

Another consistently reported feature is the vocalization of dark ɫ in a syllable coda, which seems to be the most widespread variation of l. Vocalization of syllabic, word final dark ɫ is not limited to non-native varieties of English and has been reported for RP as well (Wells, 1982: 258-259). There are consistent reports of word initial variation of the voiced dental fricative /ð/ as well as of the voiceless dental fricative /θ/. This is unsurprising, as dental fricatives vary in many L1 and L2 varieties of English. It is worth noting that much of the observed variation appears in word initial position, due to the nature of the used stimuli. Similarly, variation of the voiced postalveolar fricative /ʒ/ has been reported by different authors. These features have been reported by different researchers working with different samples of participants and therefore could be, for the time being, considered as L1 and dialect agnostic, prototypical features of Chinese English phonology.

Number	Feature	Frequency	Example
1	Absence of reduced vowels	A	considered [kɒn'sɪd.əd]
2	Vocalization of /l/	A	fool [fʊə]; successful [sək'sesfɔː]
3	Voiced and voiceless dental fricative variation	A	north [noːs]; mother ['mazə]
4	Voiced postalveolar fricative variation	A	usual ['juːjɔː]
5	Epenthetic vowels	B	next to ['nekstə tu]; cried [kə'rɑɪd]
6	Diphthong variation	B	raising ['rɑɪzɪn]; safety ['siːftɪ] ['seftɪ]
7	Deviant word stress	B	concern [ˈkənsəːn]
8	Variation of rhoticity	B	car [kɑː] [kɑr]

Table 2.2: Categorization of phonological features of Chinese English in analogy to the eWAVE categorization

There are other features that could be candidates for the prototypical category. Epenthetic vowels have been reported either in word final position or in multiple positions (see section 2.2.1 above). Some form of diphthong variation has been reported in different studies (see section 2.2.1. Deviant wordstress has been reported, ranging from stressed function words to syllable timed rhythm. There also seems to be considerable variation in rhoticity, which might be due to the fact that the language model is British English but many foreign teachers are American (Y. Jiang, 2002: 10). Table 2.2 summarizes the proposed phonological features of Chinese English, using the occurrence frequency coding used in the eWAVE (Kortmann, Lunkenheimer, & Ehret, 2020). However, the features of the varieties described in the eWAVE are mostly grammatical features, whereas table 2.2 only lists phonological features.
In order to establish prototypical grammatical features of Chinese English, a starting point could be

Number	Feature	Frequency	Example
14	No number distinction in reflexives	A	ourself for 'ourselves'
42	Object pronoun drop	A	"A: You got tickets? – B: No, Ø sold Ø already'A: Do you have tickets? B: No, I sold them already'"
43	Subject pronoun drop: referential pronouns	A	"A: You got tickets? – B: No, Ø sold Ø already'A: Do you have tickets? B: No, I sold them already'"
49	Regularization of plural formation: phonological regularization	A	"wifes, knifes, lifes, leafs"
57	Plural marking generally optional : for nouns with human referents	A	My sister-Ø are pretty girl-Ø.
58	Plural marking generally optional: for nouns with non-human referents	A	The tree-Ø don't grow very tall up there.
62	Use of zero article where StE has definite article	A	Did you get Ø mileage-claim for that trip?
63	Use of zero article where StE has indefinite article	A	"Then he thought, what about getting Ø girl [to marry] from India?"
66	Indefinite article one/wan	B	They seen one ['a'] green snake tangled round a tree
109	Perfect marker already	A	Were you there already? 'Have you been there before?
113	Loosening of sequence of tenses rule	A	I noticed the van I came in 'I noticed the van I had come in'
132	Zero past tense forms of regular verbs	A	I walk 'I walked' [but not That door bin close 'The door closed']
165	Invariant non-concord tags	A	"They had them in their hair, innit?"
173	Variant forms of dummy subject there in existential clauses	A	They is something bad wrong with her; It's a new person here 'There's a new person here';
174	Deletion of auxiliary be: before progressive	A	So you Ø always thinking about where where you go to work.
227	Inverted word order in indirect questions	A	I'm wondering what are you gonna do.
228	No inversion/no auxiliaries in wh-questions	A	What you doing? What he wants?
229	No inversion/no auxiliaries in main clause yes/no questions	A	You get the point? You liked India?

Table 2.3: Category A features appearing in both Hong Kong English and Colloquial Singaporean English

comparing two (to some degree) related varieties from eWAVE, namely Hong Kong English (Wong, 2020) and Colloquial Singaporean English (Lim & Ansaldo, 2020) with grammatical features of Chinese English reported in the literature. Table 2.3 shows features that have been marked "A – feature is pervasive or obligatory" in both Hong Kong English and Colloquial Singaporean English. Compared with the syntactic feature reported by Xu (2010, 2020), there are some similarities. Features 42 "object pronoun drop" and 43 "subject pronoun drop: referential pronouns" roughly correspond to Xu's null-subject/object utterances (NS/O). The adjacent default tense identified by Xu could be a subset of feature 132 "zero past tense forms of regular verbs". Lastly, Xu also showed that tags, particularly in questions, are used differently in Chinese English than in Inner Circle varieties, which might be related to feature 165 "invariant non-concord tags".

2.8 Conclusion

This paper has shown that even though a comprehensive description of Chinese English is not yet possible, there is a solid body of research that allows the compilation of a preliminary list of features of Chinese English. In particular, phonological features of Chinese English are well-researched. The absence of reduced vowels and thus the absence of a length contrast in vowels such as FLEECE and KIT, the vocalization of dark /ɬ/ in a syllable coda, the variation of voiced and voiceless dental fricatives, and the variation of voiced postalveolar fricatives appear to be particularly salient. In contrast, grammatical features need further investigation. Initial evidence suggests features such as object pronoun drop, subject pronoun drop for referential pronouns, and null subject/object utterances. Furthermore, variation of tags and tenses is also likely to be systematic. Discourse and pragmatic features of Chinese English can mostly be attributed to transfer from Chinese, both in terms of language and culture. Certain types of discourse, such as the ancestral hometown discourse, or assumptions about shared cultural knowledge, such as chairman Mao quotes, might not be accessible to speakers of other varieties of English. Similarly, modifier-modified information sequencing is likely to conflict with expectations of speakers of other varieties of English, without being ungrammatical. Another area that should be investigated is the influence of different Chinese L1s and dialects on pronunciation and on grammar, analyzing the eWAVE features. This opens up the possibility of establishing subnational varieties of Chinese English, analogous to work done on English varieties in multi-lingual African countries (see Schmied (1991b) for Kenyan English, for example).

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 416228727 - SFB 1410. Open Access funding enabled and organized by Projekt DEAL.

3 The Chinese English Vowel Space Across Three Provinces

This chapter is based on the article: Albrecht, S., Ivanova, M., & Schmied, J. (under review). The Chinese English Vowel Space Across Three Provinces.

Author Contributions:
SA, MI and JS contributed to the conception and design of the article. SA wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

Abstract

In this paper, a quantitative analysis of the Chinese English vowel space is presented, researching the influence of the variables gender, province, and data elicitation method (reading passages vs. sociolinguistic interview) on the vowel space of English speakers from Guangdong, Guangxi, and Yunnan. Chinese English vowels have been shown to vary in their location in the vowel space and in their lax-tense distinction. However, studies on regional varieties of Chinese English are rare. Based on manually annotated recordings of 30 participants, we performed acoustic analyses of the vowel formants and analyzed potential mergers using an individual vowel area overlap approach and a machine learning approach using support vector machines in a binary classification task. The results show that the measured lax vowel spaces have a relatively high size overlap with the tense vowel spaces (about 56%). Furthermore, we present evidence for mergers of the FLEECE-KIT and GOOSE-FOOT vowel pairs for the majority of our participants. Gender seems to have a significant influence on the vowel space in our data, while the different provinces have a much smaller influence. The findings are consistent with previous studies and demonstrate that the Chinese English vowel space shows clear and distinctive non-native characteristics.

3.1 Introduction

Vowel spaces are the entire multidimensional space in which people can produce vowels, most commonly visualized by plotting acoustic vowel polygons in the plane of the first two formant frequencies. Especially in the study of non-native or expanding circle varieties (Kachru, 1985) like Chinese English, vowel spaces are valuable for gaining interesting insights into these varieties' language systems. These insights can inform the development of language technologies such as text-to-speech (TTS) systems. In this paper, we focus on Chinese English as part of a wider research project on a non-native TTS system trained via neural networks based on authentic speech data sets. In order to build the Chinese English TTS system as a credible representation of the variety, it is crucial to research the features of Chinese English, and in particular its vowel spaces.

Quantitative research on vowel spaces started with descriptions of the native (American) English vowel space in the 1950s (Peterson & Barney, 1952) and similar descriptions have been published since

then on the English (Fox & Jacewicz, 2017; Hillenbrand, Getty, Clark, & Wheeler, 1995) and Chinese vowel space (Wiese, 1997). With increased interest in World Englishes, similar descriptions have been published for L2 varieties (see for example contributions to Mesthrie, Kortmann, & Schneider, 2008) like Chinese English (for example H. Wang & van Heuven, 2018; Q. Wang, 2023).

When discussing Chinese English, the diverse linguistic landscape of China needs to be considered. To begin with, there are languages of six major language families present in the PR China: Sino-Tibetan, Austro-Tai, Austronesian, Altaic, Austro-Asiatic, and Indo-European (M. W. Lee 1987 cited in Tang and van Heuven 2009, also see Kurpaska 2010). While Chinese dialectologists do not generally agree on how to classify the multitude of languages spoken in China, there is a general consensus that there are two major groups, the Mandarin branch, and the Southern (non Mandarin) branch. In general, the level of mutual intelligibility is considered lower in the Southern branch (Yan, 2006). A further complication in assessing the status quo is different terminology used in describing the linguistic landscape in China, or what constitutes a language or a dialect (see Kurpaska (2010) for discussion of the differences in English and Chinese terminology). Additionally, there has been a political trend to promote the use of (Beijing) Mandarin (also known as *Putonghua*) as Standard Chinese. Mandarin is the most widely spoken Chinese language throughout Chinese society, especially in education (Kurpaska, 2010; C. Lin & Jackson, 2021). The vowel systems of languages spoken in China vary considerably, with vowel inventories ranging from only three vowels to eleven vowels (Zee & Lee, 2007). The six Standard Chinese monophthongs in their characteristic triangular shape are illustrated in figure 3.1.

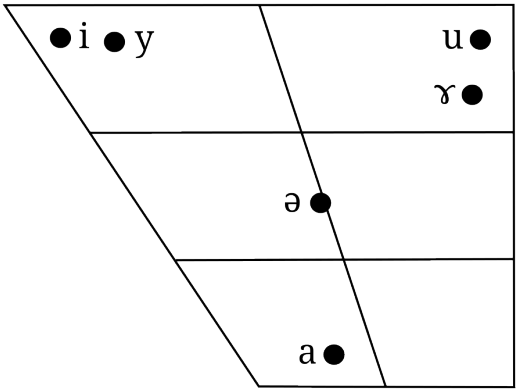


Figure 3.1: Standard Chinese monophthong inventory (from Zee & Lee, 2007: 110)

There is variation within Chinese languages such as Hakka Chinese and Shanghai Chinese (Y. Chen, 2008; A. Li, Yu, Chen, & Wang, 2004; Qiu & Liang, 2015). However, the influence of Chinese regional varieties on second language production has not been investigated, therefore, this study looks at English spoken by speakers from the provinces of Guangdong, Guangxi, and Yunnan. Of course, variation also takes place in inner circle varieties like American English, as in Californian English (A. Cheng, Faytak, & Cychosz, 2016; Fox & Jacewicz, 2017), making it difficult to speak of a unified "standard" to which Chinese English is compared. Consequently, we adopt a mostly descriptive approach, limiting comparisons with inner circle varieties.

Research on the vowel space of L2 English speakers from a Chinese background has been conducted

both in terms of perception (K. Y. Chan & Hall, 2019; Cui & van Heuven, 2011) and production (Barkana & Patel, 2020; H. C. Chen & Wang, 2011; Y. Chen & Robb, 2000; Y. Chen, Robb, Gilbert, & Lerman, 2001; Robb & Chen, 2008; H. Wang & van Heuven, 2006, 2018; Q. Wang, 2023).

The perceptual studies suggest that vowel formant frequencies, reflecting the position of the vowels in the respective vowel space, play a role in the perception of foreign accents, even for synthesized vowels (K. Y. Chan & Hall, 2019). However, vowel formant frequencies and the position of vowels in the vowel space might only play a minor role on a dialect supergroup level (Mandarin, Min, Gan) (Cui & van Heuven, 2011), suggesting other factors influencing the perception of non-native accented speech. Ultimately, it seems that vowel quality alone is not sufficient to account for the perception of a foreign accent.

The majority of previous studies on the production of Chinese English vowels have used American English as a reference. Y. Chen and Robb (2000) report a compact vowel space and less vowel separability for their Chinese English participants. Y. Chen et al. (2001) report that their Chinese English-speaking participants' vowel spaces are smaller than their American English-speaking participants and that the Chinese English speakers produce vowels with less vowel separability.

H. Wang and van Heuven (2006) show that in their data, the lax vowel area of Chinese English is very close in size to the tense vowel area. Furthermore, their individual data points were spread out rather than clustered around a centroid. Although the study is a useful account on Chinese English vowel spaces, the vowels were recorded in /h_d/ environments rather than reading passages or interviews. While this is beneficial in terms of data comparability, speech will be less natural and participants might be more aware of their pronunciation (Podesva & Zsiga, 2013: 180), which in turn might influence vowel quality, quantity, and full and reduced vowel production.

Building on earlier work presented in Y. Chen and Robb (2000), Robb and Chen (2008) report results similar to H. Wang and van Heuven (2006). They report a significantly smaller vowel space area for the vowels /i, u, ɑ/ of their Chinese English speakers and a significantly larger vowel space area for the vowels /ɪ, ʊ, ɔ/. They conclude that this suggests a "generalized lack of acoustic differentiation" between the tense and lax vowels mentioned above (Robb & Chen, 2008: 175).

H. Wang and van Heuven (2018) show that both vowel quality and duration contribute to the identification of foreign-accented English. The vowel spaces of Chinese English male and female speakers they presented are very similar in size, with the female Chinese English vowel space being slightly larger. A recent study by Q. Wang (2023) shows that the vowel quality contrast in tense and lax vowel pairs (FLEECE-KIT, GOOSE-FOOT, DRESS-TRAP) is mostly neutralized in Chinese English. The study also proposes a Chinese English vowel inventory, which will be revisited in the current article.

From the overview of the background to the current study, it becomes evident that many previous studies elicit vowels in isolated word environments (usually /h_d/) rather than in connected speech. While this approach minimizes inference from preceding or following consonants, it underestimates the difference between reading out single words, which will be read more carefully and with full vowels, and text reading or spontaneous interview speech, which will also include vowel reduction. Consequently, it might be revealing to investigate Chinese English vowel spaces with data of connected speech, collected outside a laboratory setting.

Based on the previous studies, we can expect the non-native Chinese English vowel space to have distinct features, especially the tense-lax difference seems to be little pronounced. However, there seems to be no clear consensus on the shape of the Chinese English vowel space.

Previous studies often investigated Chinese English from a Second Language Acquisition perspective,

focusing on negative language transfer from learners' native language (L1) on the foreign language (L2). The current study is not interested in the question of whether the L1 influences L2 learning (Lado, 1957; Marton, 1981) or has little to no influence (Krashen, 1982), but rather aims to provide a quantitative description of the vowel space of Chinese English speakers from three different provinces in two different speech elicitation conditions.

In particular, this study aims at a description of the Chinese English vowel space of participants from previously researched provinces of China (namely Guangxi and Yunnan, see Albrecht (2021) for a detailed review of previous studies), as well as from a province that has not been previously researched (Guangdong). To achieve this goal, we investigate the following research questions:

(RQ1) What are the vowel space features of Chinese English from Guangdong, Guangxi, and Yunnan?

(RQ2) What regional or gender differences are there in the Chinese English vowel spaces under investigation in this study?

(RQ3) Does the method of data elicitation (reading scripts vs. interviews) impact the vowel spaces?

Building on the results of previous studies, we hypothesize that the vowel space will have no clear tense-lax contrast and low vowel separability, meaning high variability in the vowel formant data and limited clustering of measurements of different vowels. Whether or not province and gender influence the vowel spaces is more difficult to predict, as there is little previous research enabling us to generate solid hypotheses. Following the basic assumption that people monitor their speech less in the interview part than the reading part of a sociolinguistic interview (Meyerhoff, 2011: 36) vowel spaces might show greater variation for the interview condition.

3.2 Methods

In order to answer our research questions, reading passages of 30 speakers of Chinese English from Guangdong, Guangxi, and Yunnan were recorded. The three province groups had five male and five female participants each. Of these 30 speakers, 29 were also interviewed, and one participant declined to be interviewed. Of the 30 participants, five (two Guangdong females, one Guangdong male, and one Yunnan male and female each) had to be excluded from the analysis due to issues with the recording. Participants were recruited at Sun Yat-sen University in China and were between 18 and 27 years old with the majority in the 18-24 range. On average, participants had studied English for an average of 11.3 years (standard deviation = 2.8) and majored in MINT subjects (19), the Humanities (6) and Business (4). The majority listed *Putonghua* as their L1 and only 10% named Cantonese or a local dialect as their L1. However, participants might have answered in what they thought would be a socially desirable way in the exclusively *Putonghua* based educational context. Participants were asked to read several short linguistic reading passages (Arthur the Rat, The North Wind and the Sun, The Boy who Cried Wolf, see Deterding, 2006a) and a longer list of sentences optimized for training neural TTS systems by ensuring sufficient phoneme coverage (Taubert et al., 2022). Participants were also interviewed using questions based on Tagliamonte (2006) and supplemented with additional questions relevant to the lives of Chinese university students. The audio files were recorded with a resolution of 16 bits per sample.

Both the linguistic reading passages and five-minute samples of the interviews were initially automatically force-aligned using the Montreal Forced Aligner (McAuliffe et al., 2017a). However, since the output of the MFA was not usable, which was largely due to the lack of fluency of our participants, ten tokens of each vowel for every speaker of the reading data and all vowels for the interview data were manually annotated in Praat (Boersma & Weenink, 2023). Using a Praat script (Albrecht, 2023c), formant frequencies at 20%, 50% and 80% vowel length were extracted. Afterwards, further analysis was carried out in R (R Core Team, 2023). The three measurements of the first three formant frequencies were averaged for every token and filtered for outliers using Hampel filtering (Hampel, 1974), which utilizes a configurable sliding window of three absolute standard deviations in this case. This method applies a unique filtering threshold for every data set, making it very versatile. The formant frequencies were then converted to the Bark scale (Traunmüller, 1990). The Bark scale was chosen since it sufficiently removes idiosyncratic variation due to biological differences between subjects while minimizing the possible loss of information in the data and without distorting the vowel space area (as a Lobanov z-score normalization would for example; see Flynn and Foulkes (2011) on different vowel normalization methods and Adank et al. (1999) for a discussion of different normalization methods and on why for certain conditions (like preserving regional accent differences) even raw frequencies seem best).

The complete data set was then split into a tense data set, containing the vowels /i/, /æ/, /ɔ/, /o/, /u/, /e/, and /ɑ/ and a lax data set, containing the vowels /ɪ/, /ɛ/, /ʌ/, and /ʊ/. Following Strange, Bohn, Nishi, and Trent (2005), we classify /æ/ as a tense vowel. Vowel space plots were generated by an R script (Albrecht, 2023b) using the dplyr (Wickham et al., 2020) and ggplot2 (Wickham, 2016) packages. To mitigate the issue of diphthongization of the vowels /e/ and /o/ (Haddican, Foulkes, Hughes, & Richards, 2013), we only used the measurement at 20% of their duration in our analysis in R. In order to further investigate potential mergers, we applied a combination of approaches. First, we tested the influence of the sociolinguistic variables under investigation (province: Guangdong / Guangxi / Yunnan; gender: male / female, elicitation: reading / interview) on the production of vowels, which were represented through the first and second formant frequencies. In order to test the differences in pronunciation of the commonly studied tense-lax pairs TRAP-DRESS, FLEECE-KIT, GOOSE-FOOT, we tested our data for normality with the Shapiro-Wilk test (Shapiro & Wilk, 1965), using the R rstatix package (Kassambara, 2023b), and for homogeneity of variances, using Levene's test (Levene, 1960). Based on the results of these tests, we applied the Kruskal-Wallis test (Kruskal & Wallis, 1952) as a non-parametric alternative to ANOVA.

Second, based on the results of the Kruskal-Wallis tests, two subsequent analyses were carried out only based on the reading data: (1) polygon area overlap and (2) a binary classification task based on a SVM learning analysis. These analyses did not consider the elicitation method because, on the one hand, the Kruskal-Wallis tests did not show large differences in terms of elicitation method, and on the other hand some interviews did not contain enough vowel tokens of a particular type to carry out the analyses. For the polygon area overlap (1), we calculated the overlap of individual vowel space, based on the approach presented in Albrecht et al. (2022). To investigate a potential merger for a given group with this method, for example, the FLEECE-KIT merger for the Guangdong female group, first the convex polygons hull of all FLEECE and KIT tokens by female speakers from Guangdong are calculated separately. Secondly, the overlap between these two polygons is calculated, resulting in a percentage, with percentages over 50% being interpreted as likely mergers. For the binary classification task (2) we followed a machine learning approach similar to H. L. Chan and Fuchs (fc.), for which we used the

support vector machine (SVM) method of the R package e1071 (D. Meyer et al., 2022) in a binary classification task. Analogously to Chan & Fuchs, we split the data sets into a training set (two-thirds of the data) and a validation set (one-third of the data) and ran the SVM analysis using a radial kernel for 50 iterations. Finally, the kappa values (κ) denoting model fit and classification accuracy were average over the 50 runs. The training and validation data contained the vowel labels as well as the vowels' F1 and F2 values. High kappa value indicate that the algorithm can distinguish the two categories of tokens well. The accuracy values represent the percentage of correctly classified tokens from the validation set. In a binary classification task, an accuracy value above 0.50 is better than chance. The results will be interpreted as follows: if the algorithm cannot distinguish between two vowels in question, they likely merge.

3.3 Results

To investigate the vowel space features of Chinese English spoken in three provinces (RQ1), we first describe the vowel spaces of our participants separated by province and gender (figure 3.2). Figure 3.3 presents the overall vowel spaces by province and for the Chinese English sample as a whole. In general, the vowel spaces appear rather similar in shape and size, with a large overlap between the tense and lax vowel spaces (mean = 56.12%, standard deviation = 2.60%).

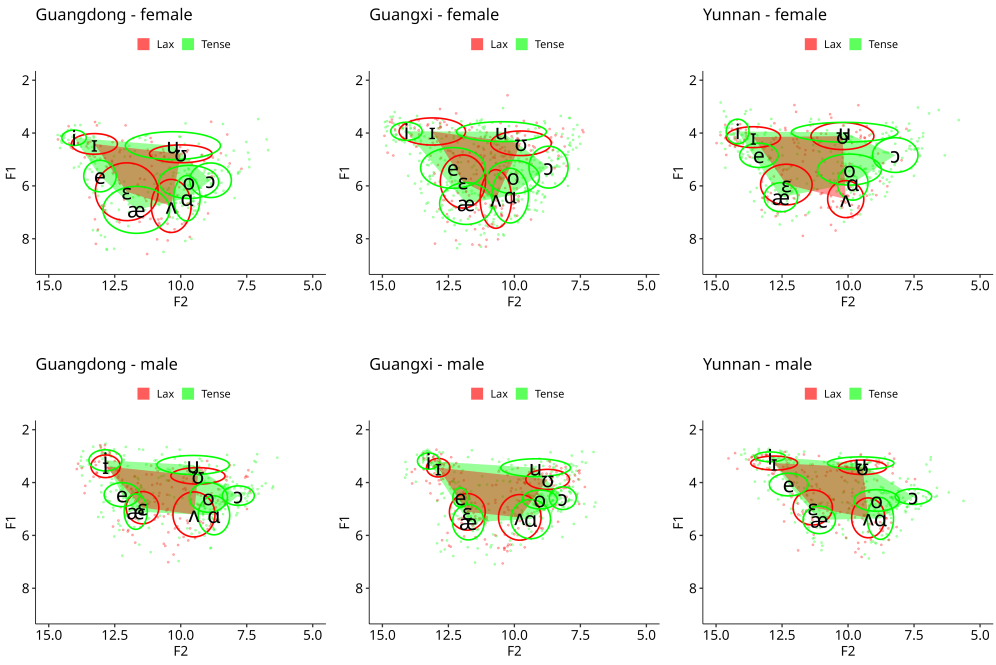


Figure 3.2: Vowel spaces of female and male participants from Guangdong, Guangxi and Yunnan in bark. Vowel means form shaded vowel space area. Error ellipses encircle one standard deviation on each axis.

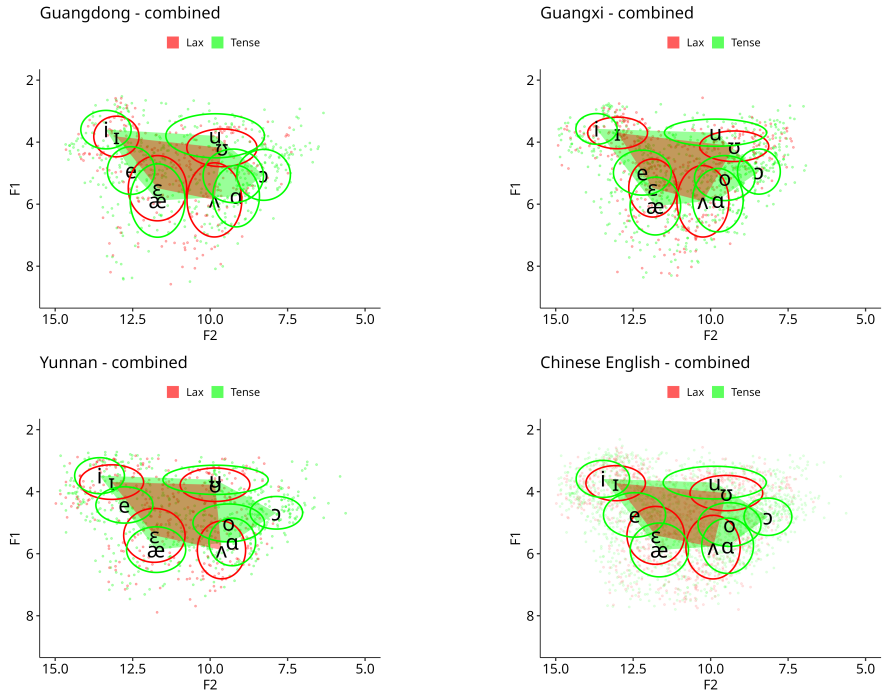


Figure 3.3: Vowel Spaces of participants from Guangdong, Guangxi, Yunnan and all the groups combined in bark. Vowel means form shaded vowel space area. Error ellipses one standard deviation on each axis.

The quantified overlap between the tense and lax vowel areas is presented in table 3.1 below. Furthermore, the CLOTH vowel is consistently very far back in our data, similar to results reported by Q. Wang (2023). The general shape and the large overlap between the tense and lax vowel area is consistent with previous accounts on Chinese English (H. Wang & van Heuven, 2006, 2018). In addition, figure 3.2 suggests that there might be mergers of the TRAP-DRESS, FLEECE-KIT, and GOOSE-FOOT vowel pairs, possibly more so for male than female participants. The combined data plots in figure 3.3 also suggest these mergers.

	Chinese English	Guangdong English	Guangxi English	Yunnan English
Combined	57.27%	54.55%	56.91%	58.02%
Female	53.38%	54.36%	50.98%	56.84%
Male	58.37%	54.35%	60.24%	57.91%

Table 3.1: Overlap percentage of tense and lax vowel space per province and gender

In order to further corroborate the existence of these mergers, we used three different approaches, the results of which will be described below. Table 3.2 shows the results of Kurskal-Wallis tests on the sociolinguistic variables gender and province, as well as the speech elicitation method (reading/interview).

Data Set	Variable	Formant	Merger	χ^2	p	Significance	η^2
ChE	elicitation	F1	TRAP-DRESS	13.54	<.001	***	0.01
			FLEECE-KIT	41.33	<.001	***	0.04
			GOOSE-FOOT	5.55	<.001	*	0.01
		F2	TRAP-DRESS	0.91	.4		0
			FLEECE-KIT	0.31	.64		0
			GOOSE-FOOT	46.99	<.001	***	0.06
	gender	F1	TRAP-DRESS	361.44	<.001	***	0.37
			FLEECE-KIT	395.77	<.001	***	0.43
			GOOSE-FOOT	296.47	<.001	***	0.4
		F2	TRAP-DRESS	63.02	<.001	***	0.06
			FLEECE-KIT	192.53	<.001	***	0.21
			GOOSE-FOOT	87.87	<.001	***	0.12
	province	F1	TRAP-DRESS	0.98	.65		0
			FLEECE-KIT	2.19	.4		0
			GOOSE-FOOT	16.64	<.001	***	0.02
		F2	TRAP-DRESS	0.04	.98		0
			FLEECE-KIT	24.41	<.001	***	0.02
			GOOSE-FOOT	5.53	.09		0.01
Guangdong	elicitation	F1	TRAP-DRESS	3.4	.09		0.01
			FLEECE-KIT	4.77	.04	*	0.01
			GOOSE-FOOT	0.12	.74		0
		F2	TRAP-DRESS	0.2	.68		0
			FLEECE-KIT	2.57	.14		0.01
			GOOSE-FOOT	24.55	<.001	***	0.11
	gender	F1	TRAP-DRESS	150.62	<.001	***	0.51
			FLEECE-KIT	121.1	<.001	***	0.46
			GOOSE-FOOT	126.84	<.001	***	0.6
		F2	TRAP-DRESS	25.13	<.001	***	0.08
			FLEECE-KIT	45.14	<.001	***	0.17
			GOOSE-FOOT	15.77	<.001	***	0.07
Guangxi	elicitation	F1	TRAP-DRESS	0.69	.47		0
			FLEECE-KIT	14.14	<.001	***	0.03
			GOOSE-FOOT	2.5	.14		0.01
		F2	TRAP-DRESS	2.62	.14		0
			FLEECE-KIT	0.28	.64		0
			GOOSE-FOOT	22.15	<.001	***	0.07
	gender	F1	TRAP-DRESS	100.18	<.001	***	0.27
			FLEECE-KIT	168.9	<.001	***	0.44
			GOOSE-FOOT	74.55	<.001	***	0.25
		F2	TRAP-DRESS	1.88	.21		0
			FLEECE-KIT	50.07	<.001	***	0.13
			GOOSE-FOOT	60.68	<.001	***	0.2
Yunnan	elicitation	F1	TRAP-DRESS	17.38	<.001	***	0.06
			FLEECE-KIT	26.11	<.001	***	0.09
			GOOSE-FOOT	7.64	.01	**	0.03
		F2	TRAP-DRESS	0.3	.64		0
			FLEECE-KIT	2.7	.14		0.01
			GOOSE-FOOT	4.92	.04	*	0.02
	gender	F1	TRAP-DRESS	112.7	<.001	***	0.38
			FLEECE-KIT	115.72	<.001	***	0.42
			GOOSE-FOOT	112.23	<.001	***	0.49
		F2	TRAP-DRESS	50.66	<.001	***	0.17
			FLEECE-KIT	92.15	<.001	***	0.34
			GOOSE-FOOT	20.17	<.001	***	0.34

Table 3.2: Results for vowel formants from Kurskal-Wallis tests. (Note: elicitation and gender $df = 1$, province $df = 2$. Notation: * significant ($p < .05$), ** very significant ($p < .01$), *** highly significant ($p < .001$). Effect sizes: $0.01 \leq \eta^2 < 0.06$ (small effect), $0.06 \leq \eta^2 < 0.14$ (moderate effect) and $\eta^2 \geq 0.14$ (large effect))

The non-parametric test can be interpreted similarly to ANOVA as it shows whether two or more groups are different. The effect sizes η^2 furthermore help interpret the significance of the p-values. In interpreting η^2, the values commonly interpreted in published literature are: $0.01 \leq \eta^2 < 0.06$ (small effect), $0.06 \leq \eta^2 < 0.14$ (moderate effect) and $\eta^2 \geq 0.14$ (large effect).

Regarding province, for the combined Chinese English data set, province only partially influences mergers. The data shows a highly significant influence on F1 for the goose-foot merger difference and a highly significant influence on F2 for the fleece-kit merger difference between the provinces. However, in both cases, the effect size is only small. Regarding gender, the results confirm the hypothesis deduced from the vowel space plots that gender influences mergers. This holds true both for the combined data sets and for the data sets of individual provinces. Furthermore, the majority of highly significant results here also have a large effect size.

Regarding the data elicitation method, the difference between reading and interview is only partially significant in the large Chinese English data set, and mostly not significant in the province data set. In all cases, it only has a small effect size, suggesting that the reading task does not influence the vowel production of our participants. Based on these results, we will discuss the vowel space overlap only based on province and gender for the read speech subset of the data. Table 3.1 lists the overlap percentage of the tense and lax vowel spaces divided by province and gender.

Zooming in on the individual tense-lax vowel pairs, table 3.3 shows the overlap of the vowel space area spanning all data points of the two vowels in each merger.

Group	Gender	TRAP-DRESS	FLEECE-KIT	GOOSE-FOOT
ChE combined		54.65%	78.28%	87.26%
ChE	female	50.39%	27.24%	79.01%
	male	52.93%	78.69%	71.18%
Guangdong	female	27.61%	32.85%	95.21%
	male	40.54%	80.15%	62.83%
Guangxi	female	30.34%	27.68%	66.14%
	male	85.04%	64.31%	45.29%
Yunnan	female	12.27%	34.79%	68.23%
	male	11.34%	39.58%	86.57%

Table 3.3: Results of the vowel area overlap analysis

The highlighted values indicate an overlap of more than 50%, which we consider an indication of a potential merger. The results show that the GOOSE-FOOT vowels merge for all subgroups except for Guangxi males. The TRAP-DRESS and FLEECE-KIT mergers seem to be a little less common across the subgroups. In the combined Chinese English data set, all three vowel pairs seem to merge, though the overlap percentage ranges from 55% to 87%. This might be somewhat surprising, especially for the TRAP-DRESS vowel, which merges in the complete ChE data set (54.65%, female 50.39%, male 52.92%) but in no subgroups apart from Guangxi male (85.04%). This is due to the fact that the analysis was carried out for all data sets separately, meaning results should be interpreted individually and are not additive. This means that the results of the Chinese English combined data sets should be interpreted separately from the individual data sets separated by province and gender. The Chinese English results therefore are neither mean nor sum of the province and gender results. The observed gender differences

based on visual inspection of the vowel space plots presented above can partially be confirmed. For both the Guangdong and the Guangxi groups, male participants seem to have two mergers, while female participants only have one, which is consistent with the plots. Both male and female participants from Yunnan seem to only produce a GOOSE-FOOT merger, which is not directly obvious from the plots, in which the corners of the tense and lax vowel space polygons seem to be very close together, and the standard deviation ellipses of FLEECE-KIT also overlap to a large extent. The difference might be due to increased variation in the individual vowel areas.

In order to provide another measure for the investigated vowel mergers, we conduct a SVM classification task, the results of which are presented in table 3.4. The value of κ indicates the strength of agreement with the tested hypothesis, which in this case is that the two vowels under investigation are sufficiently distinct based on their first two formant frequencies.

Group	Gender	TRAP-DRESS (κ / accuracy)	FLEECE-KIT (κ / accuracy)	GOOSE-FOOT (κ / accuracy)
ChE combined		0.67 / 0.84	0.26 / 0.63	0.37 / 0.67
ChE	female	0.65 / 0.83	0.35 / 0.67	0.36 / 0.68
	male	0.70 / 0.85	0.28 / 0.64	0.43 / 0.71
Guangdong	female	0.82 / 0.92	0.35 / 0.67	0.29 / 0.64
	male	0.50 / 0.75	0.16 / 0.57	0.54 / 0.77
Guangxi	female	0.57 / 0.79	0.42 / 0.71	0.49 / 0.75
	male	0.63 / 0.82	0.31 / 0.65	0.49 / 0.75
Yunnan	female	0.89 / 0.94	0.17 / 0.55	0.25 / 0.63
	male	0.84 / 0.93	0.24 / 0.61	0.14 / 0.54

Table 3.4: Results of the SVM classification analysis. Small kappa values indicate poor separability.

Table 3.5 presents the interpretation for κ values based on Landis and Koch (1977). While "clearly arbitrary [... these values] provide useful 'benchmarks'" (Landis & Koch, 1977: 165) for interpreting κ values. The accuracy measure in table 3.4 indicates the percentage of correctly classified vowels in the validation data set, meaning the sum of true positive and true negative results divided by the sum of true positive, true negative, false positive, and false negative results. The combination of κ and model accuracy provides a good estimate of the ability of the model to differentiate between the two given categories, which in this case are the vowel pairs under investigation.

κ	Strength of Agreement
<0.00	poor
0.00 - 0.20	slight
0.21 - 0.40	fair
0.41 - 0.60	moderate
0.61 - 0.80	substantial
0.81 - 1.00	almost perfect

Table 3.5: κ interpretation table (from Landis & Koch, 1977)

The highlighted values in table 3.4 are values that fall into the lower half of the κ interpretation categories presented in table 3.5. Similar to table 3.3, the results should be interpreted individually and are not additive.

The FLEECE-KIT vowel pair seems to be the most difficult for the algorithm to differentiate, with only the Guangxi female group having a κ value of over 0.40. Notably, however, the accuracy for groups with κ values below 0.4 is also low, especially for the Guangdong male and Yunnan female groups only being slightly above chance. Meanwhile, the TRAP-DRESS vowel pair seems to be relatively easily distinguishable for the algorithm, with consistently high κ and accuracy values. The GOOSE-FOOT vowel pair is also difficult to distinguish for the algorithm for around half of the subgroups and the combined Chinese English data set.

Table 3.6 summarizes the results of the vowel space overlap approach and the binary classification approach that were used to investigate potential mergers.

	TRAP-DRESS		FLEECE-KIT		GOOSE-FOOT	
	Female	Male	Female	Male	Female	Male
ChE combined	Y/N		Y/Y		Y/Y	
ChE	Y/N	Y/N	N/Y	Y/Y	Y/Y	Y/N
Guangdong	N/N	N/N	N/Y	Y/Y	Y/Y	Y/N
Guangxi	N/N	Y/N	N/N	Y/Y	Y/N	Y/N
Yunnan	N/N	N/N	N/Y	N/Y	Y/Y	Y/Y

Table 3.6: Combined indicators of mergers (overlap analysis / SVM analysis)

The highlighted cells indicate subgroups for which both methods suggest a merger. Consequently, for our Chinese English data set, both FLEECE-KIT and GOOSE-FOOT seem to merge, while TRAP-DRESS are differentiated. Looking at the participant groups divided by province and gender, every group seems to differentiate between trap and dress, but not necessarily between the other two vowel pairs. Moreover, the male groups seem to have slightly more mergers than the female groups, as four male and three female groups show mergers. The evidence for the mergers of the FLEECE-KIT and GOOSE-FOOT vowel pairs becomes even more robust when also considering data sets for which only one test suggests a merger.

Based on the findings presented above, we propose the tentative Chinese English vowel inventory shown in figure 3.4. Since both FLEECE-KIT and GOOSE-FOOT merge in our data and since the absence of vowel reduction has been reported in the Chinese English literature and is a widespread feature of non-native varieties of English[1], we suggest that only the full FLEECE and GOOSE vowels are part of the Chinese English vowel inventory. Furthermore, both vowels are also part of the Mandarin Chinese vowel inventory shown in figure 3.1. Our proposed vowel inventory differs from Q. Wang (2023) by not including the reduced forms (KIT and FOOT) for the mergers we found while maintaining dress and TRAP as separate vowels.

[1] The absence of vowel reduction seems to be a widespread feature of non-native varieties of English and has been reported, for example in Polish English (Porzuczek, 2012), Cameroon English (Simo Bobda, 2008), Nigerian English (Gut, 2008) or Jordanian Arabic English (Abu Guba, Mashaqba, Jarbou, & Al-Haj Eid, 2023).

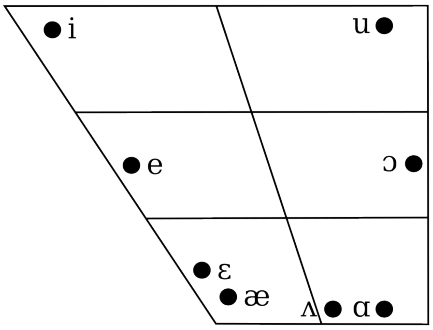


Figure 3.4: Proposed, tentative Chinese English vowel inventory

3.4 Discussion

Our phonetic analysis confirmed that the Chinese English vowel space does in fact have a consistent, unique, non-native shape and that the tense-lax distinction is not very pronounced, with likely merger of FLEECE-KIT and GOOSE-FOOT (RQ1). We furthermore showed that there are small differences in the vowel spaces by province and larger differences in terms of gender (RQ2), yet the general size and shape of the Chinese English vowel spaces across the province and gender subgroups are very similar. Furthermore, while the data elicitation method showed a significant influence in the combined Chinese English data set, the effect size was rather small. In addition, the data elicitation method did not show a significant influence on the province data set, suggesting that overall the data elicitation method does not impact the Chinese English vowel space (RQ3).

Explaining the gender differences we found in our analysis is challenging. One intuitive account would be physiological differences in vocal tracts between men and women. However, transforming raw Hertz values to the Bark scale should remove differences these differences from the data. In combination with a sufficiently large sample (see Adank et al., 1999), physiological differences can mostly be eliminated as a source of variation. The general sociolinguistic principle that female speakers use the standard variety, which in this case would be a variety that more closely resembles an inner circle variety, more than male speakers does not hold true either, as table 3.6 shows that the FLEECE-KIT merger is predominantly found in the male participant groups, as expected, the GOOSE-FOOT merger, however, is predominantly found in the female participant groups. It might be idiosyncratic variation in our data, but ideally, this would be confirmed in a future study.

Our results are in line with previous studies, such as H. Wang and van Heuven (2006, 2018), in terms of the shape of the Chinese English vowel space. While the position of the CLOTH vowel in our data differs from these two studies, Q. Wang (2023) also reports the CLOTH vowel in a similar position. Consistent with other previous studies (Barkana & Patel, 2020; Y. Chen & Robb, 2000; Robb & Chen, 2008; H. Wang & van Heuven, 2006, 2018), our data shows less vowel separability in the vowel space and large variance, meaning individual vowels are sometimes scattered all over the vowel spaces. This is consistent with similar studies on other non-native varieties of English (H. L. Chan & Fuchs, fc.) and hints at generalized features of non-native English.

The accuracy of the vowel formant measurements presented here largely depends on the accuracy of

the transcription and the alignment of transcription and audio. Initially, we planned on using reading scripts to collect our data, allowing us to largely eliminate inaccurate transcriptions as a source of inaccuracies. Previous studies testing the Montreal Forced Aligner have shown that it performs very well at word and phoneme level, at a level comparable to trained human transcribers, but with more alignment errors of above 25ms (MacKenzie & Turton, 2020; McAuliffe et al., 2017a). Combined with the large data sets analyzed in this paper and filtering outliers and invalid values this should have provided a very good database. However, in reality, this step of the research process proved to be most challenging. As it turns out, our participants often times made mistakes in reading the scripts, such as reading the present tense even though the word printed was in a past tense or generally reading different words (for example *children* has become chicken and kitchen in the recordings). In combination with a generally low level of reading fluency, which includes many false starts and lengthening of the final syllable of a word, presumably to read ahead, these issues led to "derailments" of the automatic alignment and rendered the output of the MFA unusable. Consequently, all vowels under investigation had to be manually annotated.

Our vowel space analysis is consistent across all participant groups - while there were some differences between the participant groups, the vowel spaces are generally similar in shape and size. Furthermore, the results are also consistent with data from other native and non-native varieties, such as American, British, Scottish, Nigerian, Italian, and Czech English, suggesting the absence of a clear tense-lax contrast in many non-native varieties (Albrecht & Schmied, 2023).

Another aspect that became apparent when transcribing the recordings is the fact that the vowel production of our participants might have been impacted by the grapheme-to-phoneme relationships in English. The (native) English vowels vary in their grapheme to phoneme mapping, sometimes statistically predictable, but not necessarily so (Venezky, 1995). Comparative studies on native reading have shown that English readers have more difficulties reading pseudowords than readers whose L1 has a more regular orthography (Aro & Wimmer, 2003). Research on L2 reading has shown an influence of the L1 grapheme to phoneme mapping (Ijalba & Obler, 2015), which might explain differing performance in L2 pseudo word reading tasks (M. Wang & Koda, 2007). Since our reading scripts were compiled mostly from 19th century literary texts in order to avoid copyright issues (there is no fair use in Germany, see Taubert et al. 2022 for a detailed discussion of the text selection), many uncommon words might have been challenging for our participants' acquired English grapheme to phoneme mapping and consequently at least partially explain their reading performance. However, the speaking fluency is equally low in our interview data, as are false starts and hesitations.

A facet that has not been investigated in this paper is the phonetic environment of vowels, especially on consonants that follow the vowel. Since consonants before or after a vowel affect vowel length, this might be an interesting direction of future research, adding vowel length, which usually correlates with the tense-lax categories, to get a better understanding of the Chinese English vowel system.

Furthermore, a future perception study could validate the proposed vowel inventory of Chinese English, ideally using our text-to-speech system to generate stimuli to be rated by participants. This approach would also allow creating stimuli with variation in different linguistic cues, not just vowel quality, and researching their impact on perceived foreign accentedness.

3.5 Conclusion

In this paper, we have shown that the non-native vowel spaces of English speakers from the Chinese provinces Guangdong, Guangxi, and Yunnan have a consistent shape, with a very weak tense-lax contrast and possible mergers of the FLEECE-KIT and GOOSE-FOOT vowel pairs. Furthermore, in our data, the CLOTH vowel is consistently far back in the vowel space. These findings are largely consistent with previous findings on Chinese English (H. Wang & van Heuven, 2006, 2018; Q. Wang, 2023). Based on our results, we tentatively suggest the Chinese English vowel inventory shown in figure 3.4 above.

Furthermore, we have shown that gender and to a much smaller extent province have a significant influence on the realization of tense-lax vowel pairs in our Chinese English data. The elicitation method of the data had no significant influence on the vowel realization. The tense-lax contrast is slightly less pronounced in the male participants of our study, which we showed in both the vowel plots and the subsequent analysis. Participants from Yunnan seem to have slightly less congruent tense and lax vowel space areas.

A desirable extension of our study would be a combination with data from Northern provinces, to see whether the L1 North-South divide of the Sintic language family is mirrored in L2 English. Furthermore, more precise location data of participants might help to create a foreign language variety map of China similar to the native dialect maps by Siebenhaar and Hahn (2019).

On the methodological side, we have shown that both vowel space overlap analysis and a machine learning-based SVM binary classification analysis can be useful tools in confirming possible mergers from vowel plots. Unfortunately, we could not utilize the forced alignment tools to prepare our data for analysis as planned but had to carry out the data annotation manually, which led to fewer vowel tokens being included in the analyses. Still, this highlights the need to study World Englishes in order to adapt language tools to work better with data of non-native speakers of English.

In conclusion, we have demonstrated that the Chinese English vowel space has clear non-native features that are consistent with the literature. While there is some variation across gender and the provinces under investigation here, the general characteristics of the Chinese English vowel space are consistent across the subgroups. Combined with work on other inner, outer and expanding circle varieties (Albrecht & Schmied, 2023), the results hint at a high lax-tense vowel space overlap as a common feature of non-native English varieties. However, more comparable data from other areas of China would be desirable, as our data has a clear focus on Southern China.

4 Durational Variation across inner/outer/expanding Circle Varieties of English

This chapter is based on the article: Albrecht, S., Ivanova, M., & Schmied, J. (under review). Durational Variation across inner/outer/expanding Circle Varieties of English.

Author Contributions:
SA, MI, and JS contributed to the conception and design of the article. SA wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

Abstract

This paper presents a quantitative analysis of durational variation in English varieties across Kachru's (1985) three circles, namely the inner circle (American, Northern British, Scottish, Irish English), the outer circle (Nigerian English), and the expanding circle (Czech, Italian, Chinese English). Traditionally, languages have been categorized as either stress-timed or syllable-timed, but recent research has challenged this binary classification, ultimately rejecting the notion of isochrony in language and instead shifting to investigating durational variation.

The study investigates the impact of language experience, learner background, region, and data elicitation method on durational variability in outer and expanding circle varieties, using inner-circle varieties as reference points. While inner circle varieties have consistently been considered "stress-timed", showing similar durational variation, the case of outer and expanding circle varieties is less clear. Nigerian English has been considered more "syllable-timed" than inner circle varieties and therefore should show different durational variation patterns. Similarly, Chinese English is considered "syllable-timed" in the literature. There is limited research on Italian and Czech English, but early findings suggest L1 influence.

To analyze durational variation, several durational variation metrics (DVMs) are employed, measuring vocalic and consonantal intervals. The results show that metric outcomes vary between inner, outer, and expanding circle varieties, with expanding circle varieties exhibiting more diverse results. The study suggests possible influences of articulation rate and data elicitation method on metric results but finds that these differences are not statistically significant. Overall, the research demonstrates the utility of an automated analysis pipeline using speech rhythm metrics in quantifying impressionistic classifications of English varieties in terms of their durational variation.

4.1 Introduction

Language rhythm is not a straightforward concept, as it comes with implicit assumptions on the nature of speech and the role of periodicity. There is ample evidence that speech is not periodic at the surface level, leaving the question of whether speech is at least perceived as periodic.

There have been numerous reports on the *global rhythmic* (or *timing*[1]) profiles of languages, based on the intuition that languages can be classified into rhythmic categories (Abercrombie, 1967; Lloyd James, 1935, 1940; Pike, 1945). However, factors contributing to these intuitions as well as the general applicability of such categories are not clear, and challenging to test empirically.

4.1.1 Durational variation in speech

In a first step, it is necessary to explore the different definitions of rhythm before turning to rhythm topologies, to ultimately answer the question of whether speech is rhythmic at all (Turk & Shattuck-Hufnagel, 2013). Rhythm can be defined according to different criteria:

1. Rhythm can be considered to be based on periodicity, which can either be (a) a surface phenomenon realized by some linguistic units or processes (Abercrombie, 1967); (b) motor control phenomenon realized by coupled oscillators (O'Dell & Nieminen, 1999); (c) a perception phenomenon where an aperiodic signal is perceived as periodic (Lehiste, 1977).

2. Rhythm as a result of abstract structuring of points in time, manifested by sufficiently salient linguistic elements as grouping structure (Arvaniti, 2009).

3. Rhythm can be seen as a result of a systematic relationship between segmental identity and segmental and prosodic context, and surface duration patterns (Dauer, 1983; Low et al., 2000; Ramus et al., 1999).

Based on these conceptualizations of rhythm, different typologies for classifying languages have been proposed.

The widespread intuitive categorization (Abercrombie, 1967; Pike, 1945) of languages into rhythmic categories, such as stress-timed or syllable-timed, relies on the concept of isochrony, meaning the periodic recurrence of surface elements, such as syllables or stress-delimited feet, of equal length[2]. Despite evidence contradicting this classification (Bertinetto, 1989; Roach, 1982), rhythm timing grouping is still widely used, especially in the context of second language acquisition (for example Ordin & Polyanskaya, 2015b).

A possible explanation of perceived isochrony might be inherently linked to other structural phenomena of language, such as vowel reduction or syllable structure (Dauer, 1983). The structure of English syllables with high onset and coda complexity and the fact that most unstressed syllables contain reduced vowels lead to a very pronounced durational contrast between strong and weak syllables in English. In combination with strong syllables occurring fairly often due to primary and secondary stress, the apparent stress-timed rhythm is created (L. White & Malisz, 2020). However, this "prominence gradient" (Nolan & Asu, 2009) seems especially steep in English[3], and is less pronounced in other languages (such as Dutch Cutler & Van Donselaar, 2001), while other languages like Korean lack metrically-contrasting elements altogether (Nolan & Jeon, 2014).

---

[1] Even though it is not clear whether rhythm and timing are commensurate (L. White & Malisz, 2020).

[2] Morae have been claimed as isochronous units in languages such as Japanese (Ladefoged & Johnson, 2014), however, this concept has been abandoned in recent research (Fuchs, 2023a).

[3] The application of native intuitions of speakers of Germanic languages on other languages might falsely induce stress perceptions ("stress ghosting" Tabain, Fletcher, & Butcher, 2014)

Another approach to a typology of speech rhythm is hierarchical timing, which does not assume isochrony of any unit governing timing (L. White & Malisz, 2020). Rather, relative timing dependencies between hierarchically-nested elements, such as the syllable as basic cyclical event (Fowler, 1983) and the stress-delimited foot as the smallest beat-inducing unit (Morton, Marcus, & Frankish, 1976) are used to classify languages. The relationship between these elements can be described as a linear function, with languages differing mainly in the constant term of the function (Eriksson, 1991; O'Dell & Nieminen, 1999). Building on observations from neuroscience, an approach based on coupled oscillators has been proposed (for example Ghitza, 2011; O'Dell & Nieminen, 1999) in an attempt to find a general model of speech rhythm.

The aforementioned variation in syllable structure is also linked to cross-linguistic systematic variation in articulation rate, which is defined as syllables per second excluding pauses (L. White & Malisz, 2020). While giving estimated rates for individual languages is not feasible due to too many confounding variables, the general tendency that languages with less complex syllable structure tend to have a higher articulation rate seems to hold true (L. White & Mattys, 2007a). In addition, exposure to a (foreign) language influences perceived articulation rate, with languages a listener is not proficient in being perceived as having a higher articulation rate (the "gabbling foreigner illusion" Cutler, 2012; Pfitzinger & Tamashima, 2006). At the same time, speakers with higher articulation rates are rated more native-like in their speech production (L. White & Mattys, 2007b). Consequently, some rhythmic variation might be due to variation in articulation rate (Dellwo, 2010).

4.1.1.1 Durational Variation Metrics

Starting from Dauer's (1983) study re-evaluating rhythmic typology in speech, numerous studies set out to empirically investigate speech rhythm cross-linguistically, however, often without taking an explicit stance on the definition of rhythm (see the discussion of definitions above; also see Turk & Shattuck-Hufnagel, 2013). These studies share the approach of measuring the durational variation of vocalic and consonantal intervals using various metrics and categorizing languages based on the values these metric produce.

Even though the metrics do not measure rhythm, as an investigation of surface timing features does not constitute an investigation of rhythm (Arvaniti, 2009; Cummins, 2002), they can still provide evidence of cross-linguistic variation in the "durational marking of stress contrast" (L. White & Malisz, 2020: 171). Consequently, rhythm metrics cannot reliably differentiate between variation due to different languages and different data elicitation methods, such as reading vs interview (Arvaniti, 2012b).

The various acoustically based speech rhythm metrics share the basic reasoning of measuring variation based on vocalic and consonantal intervals rather than on syllables or feet, since (1) languages differ strongly in the amount of vowel reduction and the complexity of consonant clusters (Dauer, 1983); (2) vowel duration and intensity might be used by infants to infer the rhythm class of their native language (Mehler, Dupoux, Nazzi, & Dehaene-Lambertz, 1996); and (3) syllabification rules differ by language and are contested even within languages (Low et al., 2000). On the computational side, segmentation of language into vowels and consonants is easier than syllabification or labeling feet.

Ramus et al. (1999) proposed three speech rhythm metrics that have been used in subsequent studies (Ramus, 2002; Ramus et al., 2003): ΔV, the standard deviation of vocalic intervals, ΔC, the standard deviation of consonantal intervals, and %V, the percentage of total speech duration that comprises vowel intervals.

While ΔV, ΔC, and %V only capture global variation, pairwise variability indices (PVI) try to capture global variation as well as syntagmatic contrasts within syllables by averaging durational differences of successive vocalic and consonantal intervals over an utterance (L. White & Malisz, 2020: 172). An early PVI metric by Low et al. (2000) produced different distinct values for a speaker of Singapore English and British English, confirming the categorization of the former variety as syllable-timed and the latter variety as stress-timed.

A shared weakness of these four variance-based metrics is their correlation with speech rate, meaning lower speech rates lead to higher variance (Dellwo & Wagner, 2003; L. White & Mattys, 2007a). This weakness is addressed by normalized pairwise variability indices (nPVI) that normalize interval duration by speech rate (Low et al., 2000).

$$nPVI = 100 \times (\sum_{k=1}^{m-1} |(d_k - d_{k+1})/((d_k + d_{k+1})/2)|)/(m-1) \tag{4.1}$$

Similarly, normalized variants of ΔV and ΔC, VarcoV (Ferragne & Pellegrino, 2004) and VarcoC (Dellwo & Wagner, 2003) respectively have been proposed. However, it has been shown that normalization in metrics based on consonantal intervals also removes linguistically relevant variation (Grabe & Low, 2002; L. White & Mattys, 2007a). Grabe and Low (2002) addressed this by proposing raw PVI metric for consonantal intervals.

$$rPVI = (\sum_{k=1}^{m-1} |d_k - d_{k+1}|)/(m-1) \tag{4.2}$$

Comparative studies found that normalized vocalic interval-based metrics (VarcoV, nPVI-V) in combination with %V accurately capture cross-linguistic variation and consonant cluster complexity (Loukina, Kochanski, Rosner, Keane, & Shih, 2011; Prieto, Vanrell, Astruc, Payne, & Post, 2012; L. White & Mattys, 2007a). However, cross-linguistic studies also showed the scores produced by these metrics to vary equally much between languages across assumed rhythm classes as within assumed rhythm classes (Arvaniti, 2012b; Grabe & Low, 2002; L. White & Mattys, 2007a). Therefore, no conclusions on speech rhythm typologies derived from speech rhythm metrics should be drawn, however, the variation within and across hypothesized rhythm categories aligns with perceptual studies (Arvaniti & Rodriquez, 2013; L. White, Luche, & Floccia, 2016; L. White et al., 2012).

Additionally, these metrics are susceptible to variance in utterance composition and idiosyncratic variation (Arvaniti, 2012b; Loukina et al., 2011; Prieto et al., 2012). While these issues could potentially be mitigated by large sample sizes and varied materials (L. White & Malisz, 2020), possibly in combination with automated analysis approaches (Wiget et al., 2010), the general issues of variation in metric performance due to data elicitation strategy (Arvaniti, 2012b) and inconsistent performance in distinguishing languages (Loukina et al., 2011) remain. The issue of the correlation with speech rate can also only partially be remedied by reporting speech rate along with the values of rhythm metrics.

Note that some of the experiments in studies trying to empirically confirm or deny the existence of rhythm classes or similar typologies use stimuli that are far from natural language, such as low pass filtered, re-iterant (*sasasa*) speech (Arvaniti, 2012a).

In summary, durational variation metrics can provide approximations of broader phonetic and phonotactic patterns, but come with a set of underlying assumptions about the nature of speech that have to be considered carefully (L. White & Malisz, 2020). Starting with Lehiste (1977), who

argued that the classification of languages based on isochonic speech rhythm is largely due to perceptual processes, evidence refuting the existence of rhythm classes based on isochronic intervals has mounted (Sebastiangalles 1992; Tabossi, Collina, Mazzetti, and Zoppello 2000; Zwitserlood, Schriefers, Lahiri, and Van Donselaar 1993; for a review seee L. White 2018).

4.1.1.2 Other models of prosodic speech timing

In order to account for the multitude of factors influencing speech duration patterns, including word frequencies, emotional states, and performance idiosyncrasies (L. White & Malisz, 2020) several models have been proposed. At the segmental level, in some languages voicing as well as place and manner of articulation influence both vowel and consonant durations (Klatt, 1976; Van Santen, 1992), and vowel and consonant identity can be signaled by quality and quantity contrasts (Ladefoged & Johnson, 2014).

To address these factors influencing duration patterns, two types of approaches have been proposed. In addition to a syllable's segmental composition (Van Santen & Shih, 2000), localized lengthening determines the duration of a syllable (L. White & Malisz, 2020). To account for the different lengthening and shortening effects reported (see. for example T. H. Crystal & House, 1988; Dankovičová, 1997; Fletcher, 2010; Lehiste, 1977; Oller, 1973) the prosodic timing framework (L. White, 2002, 2014) was proposed, which states that speakers use lengthening as a durational means to signal structure to listeners.

Taking a more global approach, coupled oscillator models of prosodic timing (O'Dell & Nieminen, 1999, 2009) posit that two (or more Barbosa, 2007; Malisz, O'Dell, Nieminen, & Wagner, 2017) cyclical processes are coupled and generate complex but observable surface timing patterns. Rather than isochrony or periodicity, these models assume the cyclical nature of underlying coupled units. In contrast to localized models (L. White 2014 as discussed above), these models cannot account for the location of the durational effects. While there seems to be some suggestive support of hierarchical coupling (Cummins and Port 1998; Port 2003; Schlüter 2005; but see Temperley 2009 for counter-evidence), these models are not yet widely adopted in phonetic research.

4.1.2 Acquisition of durational variation by non-native speakers of English

As languages exhibit different levels of durational variation, learners of foreign languages have to adjust accordingly in the process of foreign language learning. DVMs have been used in several studies (A. Li & Post, 2014; Mok & Dellwo, 2008; Polyanskaya & Ordin, 2019; Stockmal, Markus, & Bond, 2005; L. White & Mattys, 2007a, 2007b), often, however, under using the term (speech) rhythm. In general, vocalic metrics have proven more robust in these studies (Algethami & Hellmuth, 2023: 5), especially in differentiating between native and non-native speech.

These studies mostly focused on L1 transfer in L2 acquisition, with the assumption that if the durational variation in the L1 is similar to the L2, the latter should be easily acquired and learners should exhibit native-like durational variation, even at intermediate levels of proficiency. While this is true for some L1-L2 combinations, for example, Dutch learners were more successful at acquiring English durational variation than Spanish learners (L. White & Mattys, 2007a), it could not be the case for other language combinations, for example for German and Chinese learners, L1 influence was shown to be insufficient to explain durational variation (A. Li & Post, 2014). Both advanced German and Chinese learners approximated VarcoV values of native speakers of English, suggesting a similar developmental trajectory due to an underlying, universal mechanism (A. Li & Post, 2014). Ordin and Polyanskaya

(2015a) reach a similar conclusion from their analysis of French and German speakers of English. A similar line of research has been pursued by studies on the effect of language experience, operationalized either by the duration of residence in a country in which the target language is spoken or by grading participants according to language proficiency. Results are inconclusive at this stage, as some studies (A. Li & Post, 2014; Ordin & Polyanskaya, 2014, 2015a) found that increased language experience led to closer approximations of native durational variation, irrespective of L1 background, while others (H.-Y. Lee & Song, 2019) did not find this effect.

4.1.3 Varieties under investigation

In this study, we will compare the speech rhythm of different variety circles (Kachru, 1985): inner circle (American English (AmE), north British English (BrE), Irish English (IrE), Scottish English (ScE)), outer circle (Nigerian English (NiE)), and expanding circle (Chinese English (ChE), Czech English (CzE), Italian English (ItE)) varieties. This research thus provides a broader overview of the rhythm of speakers belonging to three English variety circles rather than a pairwise comparison of one outer/expanding circle variety with either British or American English. Including several inner circle varieties might also yield interesting results, as Scottish English has been described as more "syllable-timed" than RP (G. Brown, 2013: 52). We also include a wider variety of metrics than other comparative studies (Fuchs, 2023b),

There is a substantial body of research on Nigerian English. For a good first overview, the Nigerian English eWAVE entry proves useful (Taiwo, 2020), while more in-depth descriptions of Nigerian English phonology (Gut, 2008), prosody (Gut, 2005), and morphology and syntax (Alo & Mesthrie, 2008) are available. Due to the diverse linguistic landscape of L1s in Nigeria, subnational varieties are a very likely possibility, especially for the languages with the largest numbers of speakers, Hausa, Igbo, and Yoruba (Gut, 2005: 38). A previous analysis of Nigerian English speech rhythm has revealed it to be more syllable-timed than British English, but less so than two other outer circle varieties, namely Pakistani and Philippine English (Fuchs, 2023a).

Chinese English has traditionally been classified as syllable-timed (Deterding, 2006b: 183), an effect that can be clearly observed in multi-syllable words (He & Li, 2009: 72). Most studies on patterns of speech in Chinese English focused on the difference between tonal and stress languages in fundamental frequency (f0) patterns (G. T. Chen, 1972; Eady, 1982; Xue et al., 2002) rather than speech rhythm metrics. However, in recent years, a number of studies reported speech rhythm metric values such as normalized pairwise variability index (nPVI) values (Ding et al., 2016; A. Li & Post, 2014; S. Liu & Takeda, 2021). These studies showed that Chinese English is in fact more syllable-timed than L1 English and that speech rhythm metrics are a valid method to place languages on the stress-syllable-timing continuum. However, the perceptual effect of Chinese English being more syllable-timed might be stronger than the effect confirmed by empirical results of studies investigating language production (Mok & Dellwo, 2008).

While there is little research on Italian English in general, speech rhythm has not been researched at all. However, looking at research on Italian might provide hints. Traditionally, Italian has been described as less stress-based (Bertinetto, 1981; Ramus, 2002). Recently, it has been pointed out that this holds true for speech production, but Italian listeners do not use durational cues (i.e., stress) in speech recognition (Braun & Geiselmann, 2011).

There is equally little research on Czech English. Volín and Skarnitzl (2010) did not explicitly

address the question of the rhythm of Czech English but rather how rhythm metrics correlate with foreign-accentedness perceptual ratings of speech produced in noise. Stimuli with higher PVI but with lower %V were rated more foreign-like in noisy conditions. Given that PVI and %V have a respective positive and negative correlation with stress-timing (Grabe & Low, 2002), Czech English was perceived as more foreign in the less stress-timed conditions. Seeing that the study does not explore the production but rather the perception of Czech English rhythm, the question of the acoustic characteristics of Czech English rhythm remains open. We should thus also look at Czech rhythm as a first clue to the rhythm of Czech English. Traditionally, Czech has been described as syllable-timed (Palková, 1994), while others have classified it as stress-timed more recently (Duběda, 2004). This difficulty of categorizing Czech English as either stress-timed or syllable-timed is also reflected in an empirical analysis using speech rhythm metrics, which revealed that Czech might actually be in the middle of the speech rhythm continuum, but the metrics employed produced partially conflicting results (Dankovičová & Dellwo, 2007).

4.1.4 The current study

In conclusion, it seems that the concept of isochrony in language cannot be proven empirically and that the concept of rhythm is challenging in both finding a suitable definition in the context of language and in investigating it empirically. However, regular alternation seems to be preferred in speech production and perception (Wiese, 2022), possibly due to shared underlying resources used to process any signal with a regular temporal structure (Kotz, Ravignani, & Fitch, 2018).

Temporal structure without assuming isochrony leaves room for durational variation, which can be captured using empirical research methods. To this end, the metrics discussed above are useful tools "in their capacity to capture parameters of temporal organization of speech material" (Volín, 2017: 81). Consequently, instead of rhythm metrics, they should be referred to as duration variation metrics (DVM) and should only be used on sufficiently large samples of speech material to account for fluctuations in the values they produce. The properties of the speech material, such as speaking style, data elicitation methods, and articulation rates, ought to be disclosed transparently.

Even though single metrics cannot capture the complex nature of durational variation in language, Durational Variation Metrics produce valid results (Algethami & Hellmuth, 2023; Kinoshita & Sheppard, 2011), given the requirements stated above are met in terms of data quality and experiment design. Therefore, the current study uses DVMs to investigate durational variation in native and non-native English speech across varieties from Kachru's three circles. In particular, the following research questions are addressed:

(RQ1) How is the language background of English speakers from three variety circles reflected in their production of rhythm as expressed through durational variability?

(RQ2) How does regional durational variation within a non-native variety compare to duration variation across countries?

(RQ3) How does the data elicitation method (reading vs. interview) influence durational variability?

While we hypothesize observable differences across circles and learner backgrounds, it is difficult to predict how these will manifest in the results produced by DVMs. Regarding the variation between the

Chinese varieties in contrast to the other varieties, a stronger difference across countries (and therefore L1s) is likely. Similarly, it also likely that there is a measurable difference between the data elicitation methods, yet predictions are difficult as word choice and as a result syllable structure and complexity will likely differ between the two conditions, so will articulation rate and the results of the interplay of these factors is unclear.

4.2 Methods

To investigate the research questions presented above, seven commonly used DVMs will be calculated for our data: ΔC, ΔV, %V, VarcoV, VarcoC, nPVI-V, and rPVI-C. The reasoning for choosing a comparatively large number of metrics is twofold: (1) it allows comparing results with a wider range of previous studies; and (2) as discussed above, all metrics come with their own set of assumptions and limitations, especially the metrics based on consonantal intervals, therefore comparing results across multiple metrics might provide a fuller picture.

4.2.1 Data

The data of the inner circle varieties has been gathered from existing corpora: the American English data set is from the LJSpeech corpus, recorded by Linda Johnson of the LibriVox project, who presumably is not a professional voice actress (Ito & Johnson, 2017), the Northern British (five female and 14 male speakers), Scottish (six female and eleven male speaker) and Irish English (three male speakers) data is from the OpenSLR database, recorded via crowd-sourcing by volunteers (Demirsahin, Kjartansson, Gutkin, & Rivera, 2020). Due to the nature of the projects that these data sets have been created in, they only contain read data and no interview data.

The data of the outer and expanding circle varieties analyzed in this paper has been recorded by conducting sociolinguistic interviews using linguistic reading passages (Arthur the Rat, The North Wind and the Sun, The Boy who Cried Wolf, see Deterding 2006a for a detailed discussion of linguistic reading passages), a reading script optimized for training neural text-to-speech systems (Taubert et al., 2022), and interview questions (based on the catalog of questions in Tagliamonte, 2006).

The Nigerian English reading data set has four male and four female Yoruba speakers (mean age 47 and 43 years respectively) and two male and four female Igbo speakers (mean age 41 and 45 years respectively). The Nigerian English interview data set only has two male Yoruba speakers, aged 37.

For the Chinese English data, 30 undergraduate students were recorded at Sun Yat-sen University in Guangzhou, China. The group comprised five male and five female students from the three provinces Guangdong, Guangxi, and Yunnan. Participants were between 18 and 27 years old with the majority in the 18-24 range. On average, participants had studied English for 11.3 years (standard deviation = 2.8) and majored in MINT subjects (19), the Humanities (6), and Business (4). The majority listed Putonghua as their L1 and only 10% named Cantonese or a local dialect as their L1. However, participants might have answered in what they thought would be a socially desirable way in the exclusively Putonghua-based educational context.

The Czech English data was recorded at the University of Brno and comprises speech by three male and nine female speakers. Participants were aged between 20 and 37, with eight participants in the 20-22 range. They had studied English for an average of 14.5 years (standard deviation 4.2 years) and mostly majored in the humanities at BA level. The Italian English data was collected at the University of

Modena and has two male and seven female subjects, majoring in English at the BA level. All participants provided informed consent. In total, approximately 700,000 data points across all data sets have been analyzed.

4.2.2 Analysis

The interview data was transcribed manually and all data was automatically force-aligned using the Montreal Forced Aligner (MFA McAuliffe et al., 2017a), which performs very accurately (MacKenzie & Turton, 2020). The MFA conveniently generates Praat TextGrid files containing two interval tiers, one for words and one for phonemes. Different pre-trained models and pronunciation dictionaries were used for American English (Gorman, Howell, & Wagner, 2011; McAuliffe & Sonderegger, 2022d), UK English varieties (McAuliffe & Sonderegger, 2022c, 2022e), and Nigerian English (McAuliffe & Sonderegger, 2022a, 2022b). Czech, Italian, and Chinese English were force-aligned using the American English model, as neither pre-trained models nor pronunciation dictionaries were available. The force-aligned data was manually checked and corrected for the Chinese and Czech English samples. The remainder of the analysis was done in two steps. The data was first transformed and automatically measured in Praat (Boersma & Weenink, 2023), and then the analysis of durational variation was performed in R (Albrecht, 2023a; R Core Team, 2023).

In the first step (following the steps outlined by Fuchs, 2023a), a custom Praat script was used to generate a new interval tier that only contained vocalic and consonantal intervals, labeled "V" and "C" respectively. Pause intervals were copied to the new tier. Since some of the pre-trained MFA models use the ARPABET (for an introduction see Klautau 2001; originally proposed by Rice 1976) to represent phonetic symbols as ASCII characters, a part of the labeling of consonants and vowels was done based on ARPABET symbols and the rest based on IPA symbols. Diphthongs were transformed into one vocalic interval in the same way as monophthongs. Interval boundaries between adjacent consonantal intervals were removed, so that consonant clusters resulted in a single interval. Lastly, the duration of all intervals was measured.

In the second phase of the analysis, we computed five interval measures (ΔC, ΔV, %V, VarcoV, VarcoC) and two pairwise variability indices (nPVI, rPVI). The interval measures were calculated across all intervals, while the pairwise variability indices were calculated incrementally for subsequent pairs of adjacent intervals. These measures and indices were then plotted in charts following the common conventions of comparing the measures (namely, nPVI-V-r/PVI-C, VarcoV/VarcC).

The articulation rate of all data was computed using a Praat script by de Jong and Wempe (2009), and the phonation rates were calculated using a custom Praat script, counting all vocalic and consonantal intervals and calculating their total length.

4.3 Results

The various metrics calculated based on our data are presented in tables 4.1 and 4.2, grouped by inner/outer/expanding circle from top to bottom. The nPVI-V values range from 56 to 84 for the reading data and are slightly higher for the interview data, ranging from 64 to 73. The rPVI-C values range from 62 to 129 for the reading data and are also slightly higher for the reading data, ranging from 83 to 150. VarcoV values range from 52 to 108 for the reading data, however, the range becomes much larger for the interview data, with values ranging from 78 to 256. VarcoC values, in contrast, show a

larger range of values in reading data, with values ranging from 66 to 141, whereas values in the interview data only range from 94 to 110. The %V values have similar ranges in both elicitation conditions, with values ranging from 37 to 52 for the reading data and from 23 to 40 for the interview data.

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
NigE-Y	59.04	90.69	60.54	86.98	77.22	53.34	38.56
NigE-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 4.1: Values of all DVMs of the varieties under investigation for the read data

Just comparing these ranges suggests a possible influence of the data elicitation method on the results of the DVMs. In order to further investigate this relationship, data was tested for normality with the Shapiro-Wilk test (Shapiro & Wilk, 1965), from the R rstatix package (Kassambara, 2023b), and for homogeneity of variances, using Levene's test (Levene, 1960). Based on the results of these tests, we applied the Kruskal-Wallis test (Kruskal & Wallis, 1952) as a non-parametric alternative to ANOVA. The results are presented in table 4.3 and show that there is a slightly significant effect for VarcoV and a highly significant effect with a large effect size for %V.

Variety	nPVI-V	rPVI-C	VarcoV	VarcoC	ΔC	ΔV	%V
NigE-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 4.2: Values of all DVMs of the varieties under investigation for the interview data

As pointed out by Volín (2017), DVMs are largely dependent on articulation rate, thus reporting articulation rate (and phonation rate) is essential to allow a more nuanced interpretation of results. We follow the general convention of calculating the articulation rate after the exclusion of all pauses. The rates for the reading data presented in table 4.5 in section 4.4 on page 77 show an interesting trend. Moving from inner to outer, to expanding circle, articulation rates decrease, with the notable exception

DVM	χ^2	p value	Significance	η^2
nPVI-V	2.273	0.132		0.085
rPVI-C	2.919	0.088		0.128
VarcoV	6.313	0.012	*	0.354
VarcoC	2.273	0.132		0.085
ΔV	1.707	0.191		0.047
ΔC	2.273	0.132		0.085
%V	11	<0.001	***	0.667

Table 4.3: Results for the influence of data elicitation method of DVMs from Kurskal-Wallis tests. (Note: elicitation method df = 1. Notation: * significant (p <.05), ** very significant (p <.01), *** highly significant (p <.001). Effect sizes: $0.01 \leq \eta^2 < 0.06$ (small effect), $0.06 \leq \eta^2 < 0.14$ (moderate effect) and $\eta^2 \geq 0.14$ (large effect))

of Italian English. The Chinese English varieties show the lowest articulation rates and also, by far, the lowest phonation rates. The articulation rate trend is also found in the interview data, however, Italian English does not stand out as much but still has the highest articulation rate of all expanding circle varieties.

To further investigate the relationship between the DVMs and articulation rate, we calculated Spearman's rank correlation ϱ for the values of the metrics and the articulation rates (see Appendix). We selected Spearman's test rather than Pearson's test because our data is not normally distributed. However, the results are inconclusive. Metrics calculated based on vocalic intervals mostly show a significant negative correlation for the reading data, but not for the interview data. There seems to be no trend in the correlations for metrics calculated based on consonantal intervals.

In line with previous studies using DVMs (Arvaniti, 2009; Nolan & Asu, 2009; Ordin & Polyanskaya, 2014; L. White & Mattys, 2007a), we plotted pairs of vocalic and consonantal DVMs belonging to the same category (pairwise measures, normalized variance measures) to illustrate the variance and spread of the metric results. Figure 4.1[4] shows the plot of pairwise metrics and figure 4.2 shows the plot for the normalized variance metric results.

For the reading data in figure 4.1, the varieties from the UK are close to each other, while American English is more distant, due to its higher nPVI-V score. The Chinese varieties are quite far from all other varieties due to their high rPVI-C scores. The outer circle varieties are close to the UK varieties. For the interview data, the outer circle variety has the lowest metric scores, and the expanding circle varieties are relatively close together. Curiously, the Chinese varieties have lower rPVI-C scores, but higher nPVI-V scores than in the reading data.

For the reading data in figure 4.2, the UK varieties are very close together again, with American English being quite far from them, but on the consonantal axis this time. The outer circle varieties are relatively close to the inner circle UK varieties again. The Chinese varieties are very distant from all other varieties with much higher values in both metrics.

[4] The varieties are labeled as follows: American English, AmE; north British English, BrE; Scottish English, ScE; Irish English, IrE; Nigerian English (Yoruba), NigE-Y, Nigerian English (Igbo), NigE-I; Italian English, ItE; Czech English, CzE; Chinese English (Guangdong), GdE; Chinese English (Guangxi), GxE; Chinese English (Yunnan), YnE.

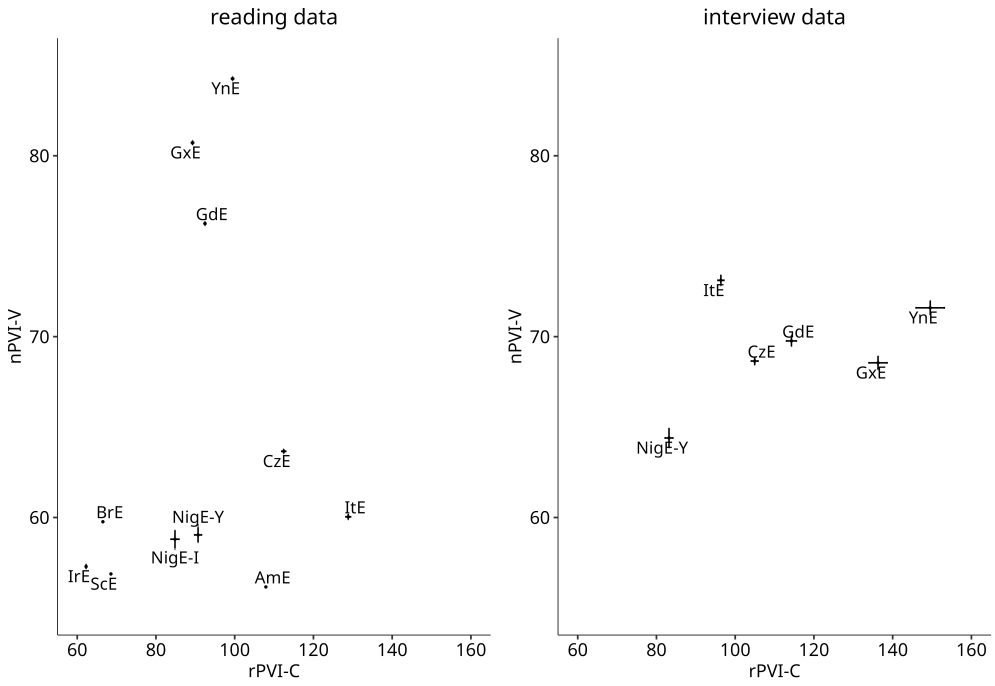


Figure 4.1: Visual comparison of reading and interview data based on nPVI-V and rPVI-C

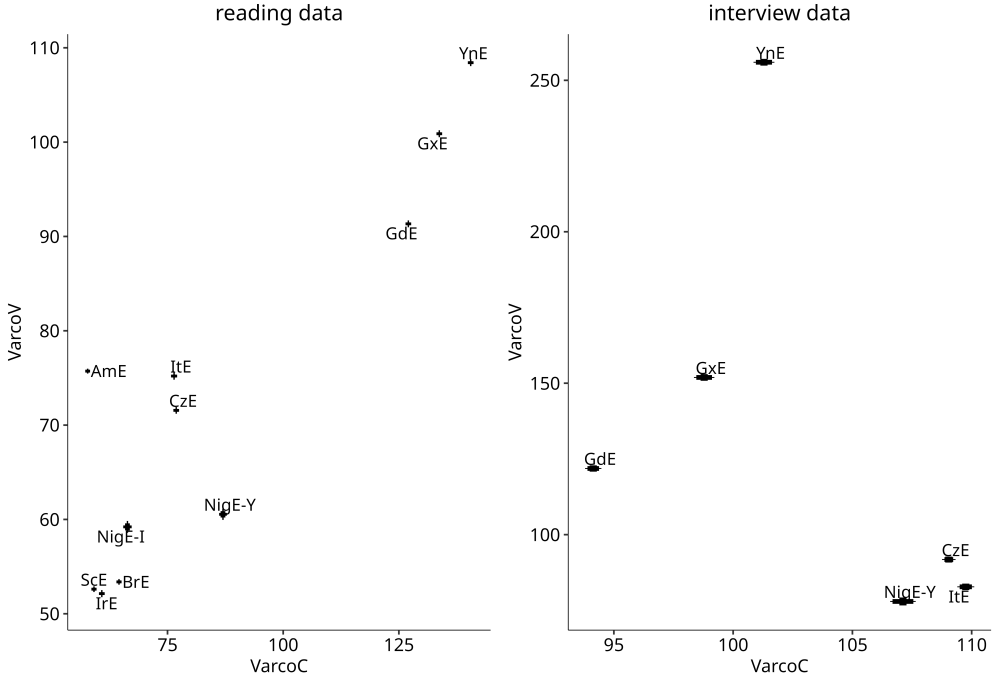


Figure 4.2: Visual comparison of reading and interview data based on VarcoV and VarcoC (Note the different scales in the subplots.)

For the reading data, Nigerian English, Czech English, and Italian English are quite close to each other, all with much higher VarcoV and slightly lower VarcoC values. The Chinese Varieties are far away from the others, with much higher VarcoC values, but lower VarcoV values.

In summary, the DVMs produce consistent results that allow the identification of tentative trends. The values of most metrics allow the identification of groups that correspond to the three circles. However, there are exceptions, as American English and Italian English show values that are different from other inner and expanding circle varieties analyzed here. The trends indicated by the values of the DVMs are also visible in the figures plotting vocalic and consonantal metrics. The most striking detail visible in the plots is the difference between the Chinese varieties and other varieties. While both the values of the metrics and the plots suggest strong differences due to the data elicitation method, the statistical analysis could only partially confirm this.

4.4 Discussion

When interpreting the results of our study, it might be insightful to compare speech rhythm metric values with the literature. Table 4.4 compares the nPVI-V values of this study with nPVI-V values of other studies of the same varieties reported by Fuchs (2023a). Our nPVI-V values of the inner and outer circle varieties are in line with the values from the literature. The values of Czech and Italian English are within a plausible range.

Variety	nPVI-V (literature)	nPVI-V (read / interviews)
AmE	51-70[5]	56
BrE	61-78[6]	60
ScE	73	57
IrE	n/a	57
NigE	55	59 / 64
ChE	40-59[7]	76-84 / 69-72
CzE	n/a	64 / 69
ItE	n/a	60 / 73

Table 4.4: Comparison of nPVI-V values with the literature

It is surprising that the results for our nPVI-V values of the three Chinese English varieties deviate strongly from the values reported in the literature. However, a qualitative investigation of the recordings revealed that the majority of our participants clearly had low fluency, which was especially evident in the interview data, despite an average 11.3 years of learning English (standard deviation 2.87). In order to quantify this impression, we calculated the average articulation rate (in syllables/second) of all our data sets. The results presented in table 4.5 reveal a trend: the inner circle varieties have the highest speech rates, and the outer circle varieties have speech rates very close to the inner circle varieties, however, the

[5] Behrman, Ferguson, Akhund, and Moeyaert (2019); Choe (2019); Ding, Lin, Wang, Wang, and Fang (2020); A. Li and Post (2014); S. Liu and Takeda (2021)

[6] Ding et al. (2016); Fuchs (2016, 2023a); Low et al. (2000); L. White and Mattys (2007a)

[7] Ding et al. (2016); A. Li and Post (2014); S. Liu and Takeda (2021)

expanding circle varieties have much lower speech rates. The notable exception here is Italian English, with an articulation rate similar to those of inner circle varieties

Variety	syllables/second	phones/second
AmE	4.28 (0.51)	10.90 (1.19)
BrE	4.38 (0.64)	8.89 (2.01)
ScE	4.45 (0.60)	9.07 (1.92)
IrE	4.64 (0.62)	9.53 (1.99)
NigE-Y	4.10 (0.28)	9.39 (1.96)
NigE-I	4.19 (0.39)	9.02 (1.36)
CzE	3.85 (0.46)	7.37 (1.26)
ItE	4.36 (0.35)	8.10 (1.03)
ChE-Guangdong	3.75 (0.30)	5.11 (1.01)
ChE-Guangxi	3.42 (0.38)	4.62 (1.00)
ChE-Yunnan	3.44 (0.37)	4.47 (0.98)

Table 4.5: Articulation rate and phonation rate of read data

Variety	syllables/second	phones/second
NigE-Y	3.97 (0.15)	4.10 (2.11)
CzE	3.75 (0.43)	5.76 (0.77)
ItE	3.85 (0.57)	5.83 (0.83)
ChE-Guangdong	3.32 (0.60)	4.94 (0.98)
ChE-Guangxi	3.09 (0.39)	4.04 (0.59)
ChE-Yunnan	3.26 (0.52)	4.11 (0.91)

Table 4.6: Articulation rate and phonation rate of interview data

This trend is also visible in the interview data in table 4.6. The Chinese English varieties have noticeably lower speech rates than the other varieties. Interestingly, the separation of outer and expanding circle varieties seems to be lost in the interview data. Curiously, the articulation rate of the Italian participants is lower than that of the Nigerian participants, in contrast to the reading data.

Another unexpected result evident in tables 4.5 and 4.6 is the fact that the phonation rate, in phones per second, for the Chinese varieties is very low for the reading data, much lower than that of any other variety, but not for the interview data. A possible reason might be that participants were much more conscious of their language in the reading condition, resulting in clear articulation but less vowel reduction. A future study may investigate this issue.

Our automated measurements depend on accurate transcriptions and correct alignment of these transcriptions with the corresponding audio. For the reading data, we aimed to achieve this accuracy through the use of standardized reading scripts that were constant across all data sets we collected ourselves. The automatic alignment produced by the Montreal Forced Aligner performs at a high level, comparable to trained human transcribers, at a word and phoneme level, but with slightly more

misalignments of above 25ms (McAuliffe et al., 2017a; McKenzie, Kitikanan, & Boriboon, 2016). While it is safe to expect some misalignments, especially when analyzing non-native speech of varieties for which no acoustic model is available, this is compensated by the ability to analyze large quantities of data.

Generally, the three circles are visible in both the DVM results and the figures, with the notable exceptions of American English and Italian English. Therefore, it seems safe to say that increased language experience, operationalized by varieties belonging to different circles, leads to durational variability more closely approximating the variability of inner circle varieties (RQ1). However, there is quite some variability within the circles as well.

The plots presented in figures 4.1 and 4.2 provide a good overview of how different the three Chinese varieties are from all other varieties. For the reading data, this effect is more pronounced than for the interview data. Furthermore, the effect can be seen in both figures, across two types of metrics. There also seems to be a gradient along which the Chinese English varieties fall, as in each figure, for both reading and interview data, the valuers of one of the metrics increase from Guangdong to Guangxi and to Yunnan. The durational variability of Yunnan English seems to have the most unique characteristics, as it has the highest values on almost all metrics while at the same time having the lowest articulation rate. Consequently, while the variation within the Chinese varieties is relatively large, the variation across countries seems to be larger (RQ2).

The observed difference of the Chinese English varieties might in fact be due to the different learning backgrounds of these learners. While the Czech English and Italian English data was elicited from English majors, the Chinese English data was elicited exclusively from non-English majors. However, quantifying this influence is difficult as the observed differences could also be attributed to L1 influence. This variable could also account for the fact that Italian English participants seemed to have different durational variation despite having a similar learning background as the Czech participants.

Lastly, both raw values and the plots show different results for the reading and the interview data. However, these impressions were largely not confirmed by statistical tests (RQ4). We only found a highly significant difference with a large effect size for the %V metric. This suggests that the difference might be due to differences in syllable structure between the two conditions. It might be that the words used in the reading passages have a more complex syllable structure and consequently the metrics produce different results. It seems that while %V is a rather simple metric, it is very robust and seems to accurately reflect underlying language variation.

4.5 Conclusion

In this paper, we have shown that varieties belonging to the inner, outer, and expanding circles produce distinctly different values on durational variability metrics. However, results have to be interpreted carefully, as they are not equally consistent across all metrics. We have furthermore shown that our Czech English and Italian English participants, despite having very similar learner backgrounds, exhibit different durational variation, which suggest L1 influence. We have furthermore shown that the durational variation exhibited by the three Chinese English varieties is very different from all other varieties investigated here.

While the values produced by the different metrics and the figures suggest an influence of the data elicitation method on the results of the metrics, this was only partially confirmed for one metric with statistical tests. While this could mean that the metric in question, %V, is the most reliable at grasping

differences between data sets, it could also mean that more variables are needed to account for the differences observed in the metrics.

It has been pointed out in previous work (Volín, 2017) that durational variability metrics are sensitive to articulation rate. However, the results of our correlation analysis are inconclusive. While the analysis showed a negative correlation between the values of the metrics and articulation rate for some metrics of the reading data, this effect was largely not confirmed for the interview data.

In conclusion, the analysis presented here has shown that durational variability metrics are useful tools to measure a surface phenomenon - durational variation. The results from their measurement and statistical analysis can be used to draw tentative conclusions on language variation and possible reasons for the measured variation. However, especially in a cross-linguistic comparison, as attempted here, controlling and accounting for all possible underlying variables influencing durational variation is challenging. Nevertheless, durational variability metrics, when paired with shared reading passages, a thorough description of participants' backgrounds, and sufficient details on the data, are valuable tools to investigate variation across different varieties and languages.

4.6 Appendix

DVM	p value (reading)	ϱ (reading)	p value (interview)	ϱ (interview)
nPVI-V	0.000 225 3	−0.9	0.92	−0.09
rPVI-C	0.5576	−0.2	0.02	−0.94
VarcoV	0.037	−0.65	0.02	−0.94
VarcoC	$<2.2\times10^{-16}$	−0.95	0.18	0.66
ΔV	$<2.2\times10^{-16}$	−0.76	0.1	−0.77
ΔC	0.67	−0.15	0.02	−0.94
%V	0.003 734	−0.82	0.8	0.14

Table 4.7: Results of Spearman's rank correlation ϱ of DVMs metric results and articulation rate for reading and interview data

5 Towards a Vowel Formant Based Quality Metric for Text-to-Speech Systems

This chapter has been published as: 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. © 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Author Contributions:
SA, RT, ST, ME, GDR, and JS contributed to the conception and design of the article. SA wrote the first draft of the manuscript, RT wrote the section on credibility and ST wrote the section on the text-to-speech system. ST implemented the TTS system. SA, RT, and ST contributed to the manuscript revision. All authors read and approved the submitted version.

Abstract

This contribution proposes an objective, vowel formant based quality metric for assessing the naturalness of monophthongs synthesized by text-to-speech (TTS) systems. This could eliminate the need for time and resource-intensive perception-based mean opinion score metrics. We show that vowel space plots can serve as a basis for developing a more comprehensive, linguistically sound quality metric for TTS systems. In addition, our metric provides detailed insights on the quality of individual monophthongs, which could help with identifying areas for further optimization in TTS systems. We use a state-of-the-art neural TTS pipeline based on Tacotron and WaveGlow, trained on the LJ Speech dataset, to generate the same audio as in our validation set and compare the two sets of audio files. Our quality metric is automatically calculated using the Montreal Forced Aligner for aligning text and audio, Praat for measuring formant values and a custom R script to calculate the vowel spaces and their overlap. Our results show that the vowel spaces of the original audio and the synthesized audio overlap to a large extent, especially for the means and the central 75% of the data, with more variation when considering all data points. Furthermore, we show that there is considerable variation in the overlap of the vowel spaces of different monophthongs.

5.1 Introduction

The research presented in this paper has been carried out as part of sub project D03 “Credibility Through Non-native Language Varieties in Conversational Pedagogical Agents” in the collaborative

research center "Hybrid Societies". The aim of this sub project is to develop a credible conversational pedagogic agent that speaks in a personalized non-native variety of English to enhance learning outcomes of non-native exchange students at German universities. The project's hypothesis is that effect that a teacher speaking in a variety close to the students own variety increases learning outcomes in sufficiently complex subject matter instruction, which has been shown in native language classrooms settings (Rey & Steib, 2013), also holds true in foreign language classroom settings and for virtual instead of human teachers.

To achieve this general goal, we modified and extended Tacotron 2 (Shen et al., 2018; Y. Wang et al., 2017), a state of the art text-to-speech (TTS) system, to work with non-native English audio input and transcripts in the International Phonetic Alphabet (IPA). To evaluate how good different non-native accents were reproduced we want to measure how close the synthesized speech matches the original recordings. Therefore, we developed an objective similarity metric and compared the vowel spaces of both human recordings and synthesized speech using our methodology described below.

Several objective and perceptual metrics have been proposed for evaluating synthesized speech quality, most of which focus on intelligibility and naturalness. In the context of evaluating synthesized speech, intelligibility is defined as the clarity of the speech, how well a listener can understand the original message. Naturalness refers to the similarity of the speech to a human voice and the general ease of listening, global stylistic consistency, regional or linguistic nuances (Sotelo et al., 2017; Taylor, 2009b).

Objective metrics used in previous research include the mel cepstral distance (MCD) (Kubichek, 1993), F0-RMSE or voice/unvoiced mismatch (VUV) (Kuo et al., 2019). Original and synthesized speech can also be aligned via dynamic time warping (Sakoe & Chiba, 1978) before applying any objective metric. Examples for subjective metrics are perception-based score tests using human evaluators like mean opinion score (MOS) (Chu & Peng, 2006), preference tests with side-by-side comparisons (Chevelu et al., 2015), intelligibility rate (IR) (Ping et al., 2018) or MUSHRA (*Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, 2015).

In general, utilizing objective metrics to evaluate synthesized speech is cheaper in terms of time and resources required compared to subjective metrics, because the latter require a large number of samples and listeners (Chevelu et al., 2015). However, subjective metrics still outperform objective ones and thus are applied in most contexts (Chevelu et al., 2015). Eventually, we plan to to combine the objective metric presented in this paper with a subjective metric as part of the credibility measurement part of the project. This will ensure a more thorough evaluation of the synthesized speech quality, and also offer insights into whether these efforts have any valid effect on the learners' perception and learning outcomes. The approach we have developed here seeks to further reduce evaluation costs by providing a novel objective metric capable of estimating the naturalness of monophthongs.

Our main contributions are:

- We developed a vowel formant-based quality metric that is suitable for evaluating the naturalness of monophthongs.

- We show that a linguistically motivated metric allows detailed insights in the quality of synthesized audio and can give directions for further optimization.

5.2 Method

We trained a TTS model on a large single speaker dataset, from which we held back a validation set of 30 minutes for calculating our metric. After training, we synthesized each sample in the validation set, aligned the original and inferred speech according to their phonemic transcriptions, and calculated our metrics for all monophthongs in the two sets (original and inferred validation set).

5.2.1 Dataset

We used the LJ Speech dataset (Ito & Johnson, 2017), which is licensed as Public Domain (CC0 1.0) and consists of recorded reading passages by a female speaker of American English. The reading texts were taken from seven public domain non-fiction books, which were published between 1884 and 1964. The audio files are of high quality, with a sampling rate of 22050 Hz and 16 bits per sample. We chose this dataset because of its properties and because it is frequently used in neural TTS applications.

We prepared the transcription by normalizing the audio files and transforming the texts into phonemes, represented by ARPAbet symbols (Klautau, 2001). This step was carried out by using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017a) and its bundled pronunciation dictionary. In the case of out-of-vocabulary (OOV) words, the phonemes were estimated using the grapheme-to-phoneme Python module g2pE (Park & Kim, 2019). Our final corpus contained 83 different symbols, consisting of 14 punctuation marks (␣`!"'(),-.:;?[]`), 30 monophthongs with stress markers[1] (`AA, AE, AH, AO, EH, ER, IH, IY, UH, UW`), 15 diphthongs with stress markers (`AW, AY, EY, OW, OY`) and 24 consonants (`B, CH, D, DH, F, G, HH, JH, K, L, M, N, NG, P, R, S, SH, T, TH, V, W, Y, Z, ZH`).

We extracted 30 minutes (470 audio files) from the entire corpus as a validation set for monitoring our Tacotron training progress and for the calculation of our quality metric. In the validation set, we tried to achieve a uniform distribution of monophthongs to ensure a sufficient amount of measurements of each of them. To do this, we minimized the Kullback-Leibler divergence (KLD) (Kullback & Leibler, 1951) between the distribution of all monophthongs in the validation set and the uniform distribution in the set of all monophthongs. The remainder of the corpus was used for training our TTS system and included about 23 hours and 25 minutes (12630 audio files).

Measurements were taken from the validation sets and a synthesized version of the same sentences, resulting in 5,494 data points for the original validation set and 5,460 data points for the synthesized validation set respectively. After filtering outliers, 5,064 data points for the original validation set and 5,563 data points for the synthesized validation set were retained.

figure 5.1 shows the distribution of monophthongs, with all forms of stress grouped together, meaning that, *AA* represents the unstressed monophthong *AA0*, the monophthong with primary stress *AA1*, and the monophthong with secondary stress *AA2*. The most frequent monophthong was *AH*, which accounted for 9.41% and 5.13% of all symbols in the training set and validation set respectively. The least frequent monophthong was *UH* with 0.22% and 0.57% respectively. The length of the bars indicates that our method results in a more even distribution of monophthongs in the validation set compared to the training set.
A breakdown of the distribution of all monophthongs in the validation set is shown in figure 5.2.

[1] In ARPAbet, stress is marked by adding 0 to unstressed phonemes, 1 to phonemes with primary stress and 2 to phonemes with secondary stress (e.g. AA0, AE1, AO2).

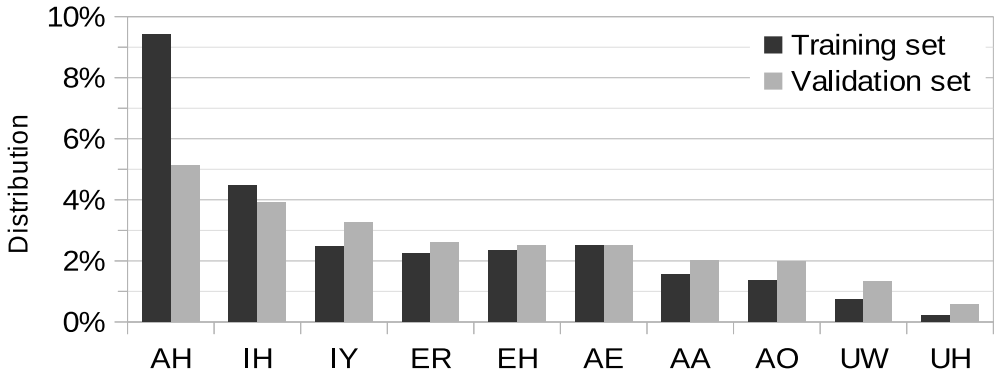


Figure 5.1: Distribution of monophthongs in training and validation set

It can be seen that vowels with secondary stress are very rare, as they would only be found in longer multisyllable words. The monophthongs *AH*, *IH*, *IY*, and *ER* occur most often in unstressed form, which is not surprising, as all of them, except *IY*, are lax vowels. All other monophthongs occur mostly with primary stress. In total, the most frequent monophthong *AH* occurred nine times more frequent than the least frequent one *UH*.

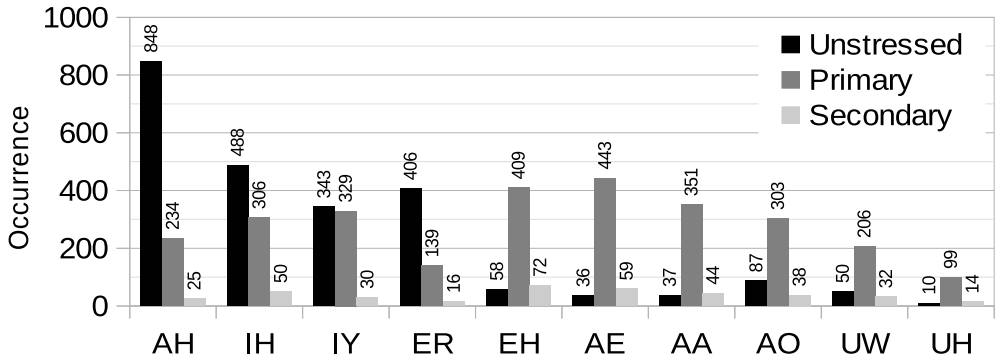


Figure 5.2: Occurrences of monophthongs in the validation set

5.2.2 Models

Our TTS system has two components: Tacotron (Shen et al., 2018; Y. Wang et al., 2017), a neural model that predicts mel spectrograms from sequences of input symbols, and WaveGlow (Prenger et al., 2019), a neural vocoder that synthesizes these mel spectrograms into audio signals. Both models are state of the art, publicly available (*Tacotron 2 - PyTorch implementation with faster-than-realtime inference*, 2018; *WaveGlow: a Flow-based Generative Network for Speech Synthesis*, 2018) and BSD-3 licensed, making them very suitable for our use case. We based our code on two public repositories, modified it for our needs and published it under an MIT license on GitHub[2][3].

[2] `https://github.com/stefantaubert/tacotron`

[3] `https://github.com/stefantaubert/waveglow`

We trained Tacotron for 500 epochs with a batch size of 64 and a learning rate of 0.001. The training took about 100 hours and was done on Ubuntu 20.04 LTS using an Nvidia Titan RTX graphics card (with CUDA 11.6) and an AMD Ryzen 7 3800X processor. For WaveGlow we used version three of a pretrained model provided by Nvidia (*WaveGlow LJS 256 Channels*, 2019), which they also trained on the LJ Speech dataset. The original and synthesized audio files from the validation set are available on GitHub[4].

5.2.3 Vowel Formant Based TTS Quality Metric

To assess the quality of the output of our neural TTS system described above, we decided to implement a linguistically motivated quality metric in addition to the built-in metric that visually compares mel spectrograms. Speech is typically said to carry three types of information: (1) linguistic information, or the message; (2) sociolinguistic information, or information on the background of the speaker; and (3) personal information, which enables speaker identification (Ladefoged & Broadbent, 1957). Acoustic analysis of vowel is especially relevant to identify variation due to sociolinguistic variables, as this variation usually occurs below the level of consciousness (Di Paolo, Yaeger-Dror, & Wassink, 2011), yet vowels carry sociolinguistic information allow identification of speakers' sex (Whiteside, 1998), age (Gahl & Baayen, 2019), regional background (Loakes, Hajek, & Fletcher, 2017) and social class (Watt, 2000). Consequently, since preserving sociolinguistic information is a central feature of our neural TTS system, developing a linguistic quality metric around vowels seemed appropriate.

The pilot study of our quality metric presented here is based on the LJ Speech dataset as mentioned above. Pairs of audio files (original recordings and synthesized) and transcripts were forced aligned using the MFA, producing TextGrid files containing individual, timestamped phonemes. These files, together with the LJ Speech audio files, were used to measure the first three formant frequencies at 20%, 50% and 80% duration of all monophthong vowels using a Praat (Boersma & Weenink, 2023) script[5], resulting in a total of 299,689 vowels measured. Further analysis was carried out in R (R Core Team, 2023), using the phonR (McCloy, 2016), tidyverse (Wickham et al., 2019), dplyr (Wickham et al., 2020) and sf (Pebesma, 2018) packages[6]. First, the data was cleaned of missing values. Then, means of the formant frequencies at the three temporal points were calculated. In the next steps, outliers were filtered using sliding window of three median absolute deviations (Hampel, 1974). Lastly, the data was Lovanov normalized (Lobanov, 1971), since this normalization algorithm generally performs very well in sociolinguistic applications. For a detailed discussion of normalization algorithms for sociolinguistic applications see (Albrecht, fc.). From the means of the normalized vowel measurements, the vowel space was computed by finding the hull of a convex polygon. Finally, the intersection of the vowel space polygons of the original recordings and the synthesized audio were computed. The percentage of the overlap of the two is our proposed vowel quality metric.

5.3 Results

The results of our analysis are illustrated by the figures and tables below. table 5.1 shows the overlap percentages of the vowel spaces illustrated in figure 5.3. In the illustrations, means refers to the vowel

[4] `https://github.com/stefantaubert/CIVEMSA-2022`
[5] Available at: `https://mytuc.org/lnvn`
[6] Script available at: `https://mytuc.org/mthv`

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

Table 5.1: Vowel Space Overlap

space described by a convex polygon covering the mean values calculated for each phoneme. The polygons of the central 75% of all data points have been calculated by finding a convex polygon that covers the middle 75% of the data points on the x-axis and the y-axis.

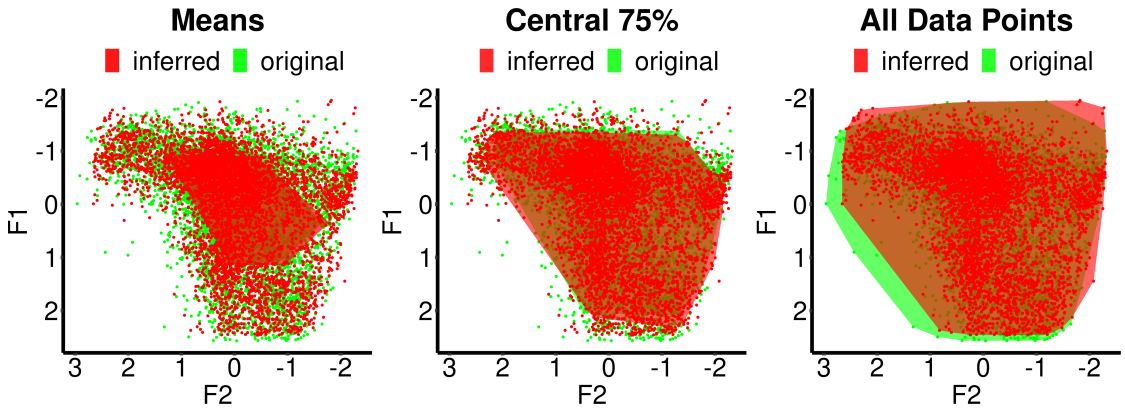


Figure 5.3: LJ Speech Vowel Space, plotted as F1 - F2 space (Lobanov Normalized, Hampel Filtered)

The results show that both the vowel space calculated from phoneme averages and the vowel space calculated from the central 75% of data points overlap greatly at 93.20% and 97.70% respectively. The vowel space covering all data points exhibit a slightly lower overlap percentage at 91.50%.

Looking at the vowel spaces of individual vowels, the results differ significantly, as there is substantial variability on the overlap percentages both between the results for the central 75% of the data points and all data points, as well as between vowels in the two categories.

While the overlap calculated across the whole vowel space exhibit greater overlap of the average and central 75% of the data compared to all data points, this trend is reversed in some of the individual phonemes, most strikingly in *AO* and *EH*.

5.4 Discussion

A limitation of the automated vowel formant measurement approach is that the forced alignment is not 100% accurate. However, the MFA generally performs well in terms of word-level and phone alignment, at a level comparable to trained human transcribers, but with more alignment errors of above 25ms (MacKenzie & Turton, 2020; McAuliffe et al., 2017a). We argue that this limitation is offset by the large amount of data we analyze and by filtering outliers. Another issue related to the forced aligner is that it depends on the pronunciation dictionary that is bundled with the software and thus, the quality aligned transcriptions depends on the quality of the dictionary. This might introduce inaccurate results for out-of-vocabulary words or words with more than one way of pronouncing them. This, in

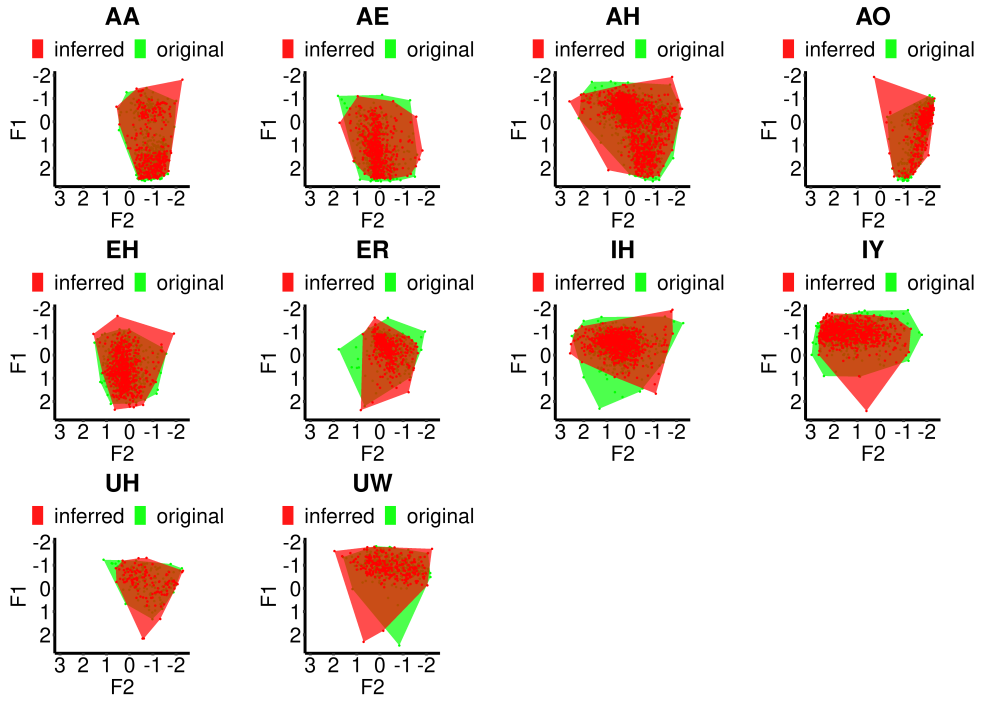


Figure 5.4: LJ Speech Phoneme Space All Data Points (Lobanov Normalized, Hampel Filtered)

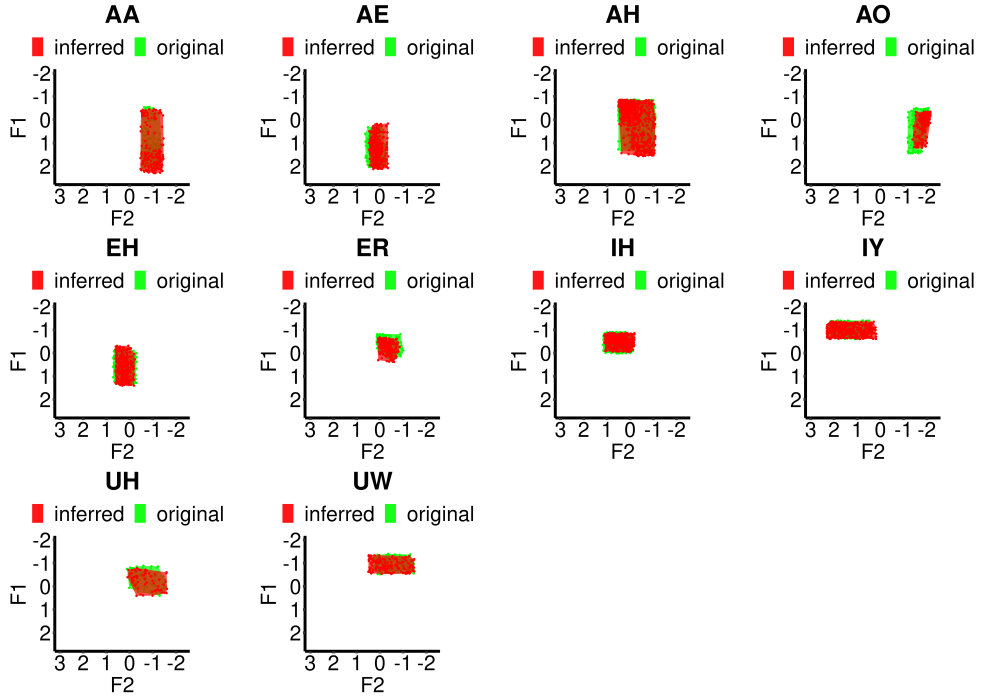


Figure 5.5: LJ Speech Phoneme Space Central 75% of All Data Points (Lobanov Normalized, Hampel Filtered)

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%

Table 5.2: Phoneme Space Overlap

turn, could also impact the measurement of credibility and subjective metrics. Especially, since cultural difference (Manikutty, Anuradha, & Hansen, 2007) and perception of accents (Cargile & Giles, 1997; Lev-Ari & Keysar, 2010) can be very important factors influencing these variables.

Furthermore, only monophthongs are measured in our automated quality metric, excluding diphthongs. This is due to the nature of diphthongs and the resulting difference in measuring them. Since diphthongs can be considered a combination of two monophthongs, plotting vowel formants in at 20%, 50% and 80% vowel duration results in a v-shaped trajectory. While this method of plotting diphthongs is widely used in acoustic phonetics, the creation of vowel space polygons and calculating their overlap as presented in this paper would be less straightforward and thus has not been implemented. Further research is needed into the feasibility of including similar plots and overlap measurements for diphthongs.

Another aspect that has not yet been considered in this study is vowel stress. As shown in figure 5.2, the validation set contain vowels in both stressed (primary or secondary stress) and unstressed form. Currently, all vowels are analyzed together, irrespective of stress. It might be worthwhile to investigate whether separate measurements for the different stress configurations can reveal additional insights into the quality of speech synthesis.

While our vowel quality metric can be calculated at different stages in the process of training a neural network, there is a certain threshold of training duration that has to be reached first. This is due to the forced alignment failing for synthesized audio from early stages of neural network training.

It is also important to keep in mind that we measured the output of our complete TTS pipeline, meaning the monophthongs in the synthesized speech generated by WaveGlow from mel spectrograms generated by Tacotron. The reasons for this are twofold. Firstly, we developed the metric described here with a specific application in mind, which requires audio files rather than mel spectrograms. Secondly, the alignment with the MFA, which is necessary to determine the position of the monophthongs, is only possible on audio files and not on mel spectrograms. Additional variability is introduced because of this combination of two neural networks, which is acceptable within the constraints of our application, however, theoretically, the formants could also be measured from mel spectrograms.

In future research, it would furthermore be interesting to investigate how our objective metric performs in comparison to subjective metrics like the frequently used mean opinion score. This could

also reveal correlations between our objective metric and more time and labor intensive subjective, perception based metrics employing human listeners. Ultimately, should our metric prove reliable, this could reduce the cost of evaluating the quality of TTS systems.

5.5 Conclusion

The vowel formant-based quality metric for monophthongs proposed in this paper can serve as a basis for developing a more comprehensive, linguistically sound quality metric for TTS systems. We have shown that vowel space plots for all monophthongs provide a good estimate of the quality of synthesized speech, with overlap percentages of 91.50% to 97.70% for the whole vowel space and 56% to 98% for individual vowels (see table 5.1 and table 5.2). Furthermore, comparing the vowel spaces of individual vowels can provide insights into how well certain vowels can be synthesized, allowing targeted optimization of training.

However, more research is needed to construct a linguistically sound quality metric for TTS systems that goes beyond monophthongs and incorporates diphthongs, consonants and other optimizations outlined above. Ultimately, it would be desirable to create a metric that could eliminate the need for participant based perceptual evaluation of TTS systems.

6 Competent but Cold: Learning with a Pedagogical Agent Speaking Non-Native Varieties

This chapter is based on the article: Ivanova, M., Albrecht, S., & Schmied, J. (under review). Competent but Cold: Learning with a Pedagogical Agent Speaking Non-Native Varieties.

Author Contributions:
MI, SA, and JS contributed to the conception and design of the article. MI wrote the first draft of the manuscript. SA implemented the survey. All authors contributed to the manuscript revision, read, and approved the submitted version.

Abstract

English is a common medium of instruction of pedagogical agents, yet most solutions incorporate only native varieties like Standard American English. Thus non-native English students perceive a difference between the language of the agent and the language of their personal sociolinguistic environment. However, it is unclear whether this leads to learning disadvantages, as there is a lack of research on the perception of non-native English speakers of their own variety. Some evidence suggests that it is beneficial when educational content is personalized in the learner's language variety. Thus, designing an agent as a credible speaker of non-native Englishes by mirroring the users' familiar sociolinguistic context has the potential to facilitate learning. We tested this hypothesis with a text-to-speech system that can be trained to speak in different varieties and can be applied in pedagogical scenarios with English as an instructional language. The agent was customized to speak Italian, Czech, and American English and was incorporated in a Linguistics lecture scenario testing learners from Italy and the Czech Republic. The results did not show an influence of the agent's variety on learning outcomes. The participants rated the speaker relatively neutral, but more favorably on COMPETENCE characteristics than on WARMTH characteristics. While the results indicate that the narration may need to be made more engaging, the current study provides first insights into the perceptions of non-native students' own language and recommendations for customizing credible, learning-facilitating pedagogical agents. Such efforts in diversifying language technologies have the potential to benefit non-native English speakers in their individual efforts to take active control and appropriate their individual learning in English.

6.1 Introduction

English has become a widespread medium of instruction in educational institutions like universities as well as in online courses and tutorials. The new technological developments in text-to-speech (TTS) and pedagogical agent (PA) solutions allow the rapid synchronization of course material, making the preparation of materials more time-efficient and easily adaptable. With the rise of the use of English in education, non-native English speakers have access to more resources. Yet, they are often disadvantaged because the language of instruction is usually different from the variety they have been exposed to

in their home country, i.e. their own variety. Karakaş (2017) examined 50 TTS products and found that they mostly featured American English (AmE) and British English (BrE), with few exceptions like Australian and Indian English. There are no known solutions for non-native speakers of English from expanding circle varieties (Kachru, 1992b) such as Italian English (ItE). While people can indeed adapt to unfamiliar varieties (Cristia et al., 2012), this takes exposure time, which could be used more efficiently to learn more information on the subject matter with less effort. Our hypothesis is that customizing a PA's voice to mirror the sociolinguistic variety of the learners has the potential to improve learning outcomes, as learners will have more perceptual and learning benefits from a familiar variety like their own.

This hypothesis is based on evidence of the advantages of learning from teachers with a familiar accent. Investigating the personalization effect, Rey and Steib (2013) found that Austrian German was more beneficial than Standard German (Rey & Steib, 2013). Listeners who heard the lecture in their local dialect scored higher in retention than those who listened to it in standard German, however, this was not the case for transfer (Rey & Steib, 2013: 2026). In addition to familiarity, language attitudes also play a role in learning. While Ahn and Moore (2011) did not find a negative effect of differently accented voice on learning outcomes, they found that learning outcomes were influenced by attitudes towards the presented varieties. Namely, only students who shared negative attitudes towards Asian English accents prior to the experiment also performed worse when presented with input produced in Korean English (Ahn & Moore, 2011).

Looking at synthesized language varieties, in an experiment with robots speaking standard Arabic and a local dialect, when the robots showed high knowledge and rhetorical ability, the robot speaking the participant's own local dialect was preferred (Andrist et al., 2015: 163). In the New Zealand context, a robot speaking New Zealand English was preferred, as it was rated more natural than the robot speaking American English (Tamagawa et al., 2011: 2011). Irish children were also more favorable towards the UK than the US voice (Sandygulova & O'Hare, 2015). Still, the evidence on the effectiveness of pedagogical agents (PAs) is mixed. In a review of 50 papers, only 12 show significant impacts of the use of pedagogical agents and of them, 76% show significant impacts on student learning outcomes (Martha & Santoso, 2019). Therefore, PAs and their application in foreign language-medium instruction is a promising research field.

It should be underlined that the current study does not investigate the teaching of English but teaching in English. In the context of the teaching of English, accent can interfere with listening exercise outcomes. For instance, Major, Fitzmaurice, Bunta, and Balasubramanian (2002) have demonstrated that while native speakers of Spanish scored significantly higher in a listening task when listening to Spanish-accented English, native speakers of Chinese scored significantly lower when listening to Chinese-accented English Major et al. (2002). Regarding the teaching in English of subject matter like Linguistics or Biology to non-native speakers, there is a research gap that we aim to address in the context of Italian and Czech English.

Due to the many different definitions of PAs, the type of agent investigated in the current study needs to be specified. Erickson (1997) has provided two definitions of a PA: 1) a (semi-)autonomous program with adaptive functionality or 2) relative to the agent metaphor, a program with animate characteristics (Erickson, 1997). While agents can vary in forms between text, voice, animation (2D or 3D), and in terms of embodiment (Martha & Santoso, 2019), they are often regarded as anthropomorphic characters (Do, Akter, Choudhary, Azevedo, & McMahan, 2022). In our study, the PA is closer to Erickson (1997)'s second definition of the agent metaphor, though it only has an authentic anthropomorphic

voice without a human visual representation. This differs from the many embodied agents in the literature featuring an animated character (see Dai, Jung, Postma, & Louwerse, 2022). Note that the PA employed is also scripted, i.e. it requires a text, which is then synthesized. The agent is only an information source and does not interact with the users (Schroeder & Gotch, 2015: 187). To sum up, our PA is a TTS system which can be applied in pedagogical scenarios.

In addition to the learning outcomes with the PA, we focus on the users' evaluation of the speaker. Both conversational and pedagogical agents are perceived as social partners (Louwerse et al., 2009; Mayer, Sobko, & Mautone, 2003) and as such, they are attributed certain characteristics based on the social and linguistic cues they send. Both designers and users of agents become constructors of vocal agents' identities, even when voice is the only modality (Seaborn, Miyake, Pennefather, & Otake-Matsuura, 2022: 30). Regardless of whether personality is a relevant factor in the robot's task, humans ascribe personality traits to it (Nass & Lee, 2001b). Agents should therefore be constructed to evoke a favorable social perception. One major factor contributing to this perception is voice, and the sociophonetic cues and features of this voice can play a decisive role when designing speech systems, as also argued by Sutton et al. (2019). Another important characteristic of the PA is credibility, which can also be influenced by sociolinguistic factors such as accent in addition to other factors like content credibility (Andrist et al., 2015; Myers & Martin, 2018). Our PA thus aims to be perceived as credible through an authentic representation of English varieties. In addition, we are interested in the attribution of characteristics in the realm of warmth and competence, i.e. whether the users perceive the agent as a likable and competent figure that is easy to understand. These categories have been successfully explored in previous research on language variety attitudes (e.g. McKenzie et al., 2016).

The current study explores how a PA speaking the users' non-native variety impacts the learning outcomes of non-native English speakers from two different cultures, Italy and the Czech Republic. The experiment simulates a learning scenario involving a lecture in Linguistics, followed by evaluation of the speaker as well as retention and transfer learning questions. The findings can be applied in the design of customizable PAs that can be adapted to the user's variety and thus facilitate learning of input produced in the language variety speakers are most familiar with - the variety spoken in their sociolinguistic environment.

6.2 Background

The design of PAs is led by several thoroughly researched effects and principles such as the social agency, cognitive effort, personalization, and voice principles (Mayer, 2014b; Mayer, Sobko, & Mautone, 2003; R. Moreno, 2012). Social agency theory argues for the importance of social cues in multimedia learning (Atkinson, Mayer, & Merrill, 2005; Mayer, Sobko, & Mautone, 2003: 419). Cognitive effort theory suggests that students work harder in a cognitively demanding situation (Mayer, Sobko, & Mautone, 2003: 421).

The personalization principle supports the use of a more conversational instead of a formal style (Mayer, 2014b: 348). The voice principle gives importance to the voice of the system, originally supporting the use of a human voice (Mayer, 2014b: 357) but later studies provide more support for computer voices (S. D. Craig & Schroeder, 2017, 2019). We will focus on the ways these principles inform the conceptualization of PAs.

6.2.1 Personalization principle

The conversational style leading the personalization principle has often been realized in the use of personal pronouns, more conversational instead of personal formulations, and direct comments to the reader (Reichelt, Kämmerer, Niegemann, & Zander, 2014). Still, studies have shown that the personalization effect is not universally advantageous but is culture-dependent. In contrast to American students, Czech students have been found to prefer more direct rather than polite messages (Brom, Hannemann, Stárková, Bromová, & Děchtěrenko, 2017).

Similarly, in an experiment with Chinese students, conversational style supported the retention but not the transfer of knowledge and also led to more pressure (L. Lin, Ginns, Wang, & Zhang, 2020: 9). Meanwhile, an "everyday" style was beneficial for German (Reichelt et al., 2014: 207) and Turkish test takers (Kartal, 2010). The attitude towards politeness can also differ within subgroups, as Lawrence, Crocker, and Blanton (2011) demonstrated for stigmatized groups' rating of positive feedback as impolite. Therefore, decisions on the application of the personalization principle should be made with the cultural and social context in mind.

6.2.2 Voice principle

Agent's voice has been found to be more important than its physical presence on the screen (Mayer, Sobko, & Mautone, 2003: 811). In Mayer, Sobko, and Mautone's 2003 study, accented voice resulted in the same retention outcomes as the nonaccented voice, but the nonaccented voice led to better transfer. In addition, the human recording outperformed the machine synthesis in retention, transfer, and subjective characteristics like dynamics and attractiveness (Mayer, Sobko, & Mautone, 2003: 422-423). Atkinson et al. (2005: 136) also found a voice effect in favor of the human voice vs the machine voice.Do et al. (2022) found an interaction effect between learner gender and synthetic speech accent for retention scores, as females performed worse under the unfamiliar accent (Indian English) condition compared to the familiar (American English (AmE)). They also noted that females rated the human-likeness of the unfamiliar accent lower (Do et al., 2022).

However, other recent studies showed that with improved voice technologies, the performance of machine voices has become similar to or better than human voices (S. D. Craig & Schroeder, 2017). S. D. Craig and Schroeder argue that "the voice effect is a by-product of technological limitations rather than a binding, persistent limitation an instructional designer should be concerned with" (S. D. Craig & Schroeder, 2019: 1537). Indeed, it has been reported that TTS can be perceived as similarly likable and trustworthy as a human (Abdulrahman & Richards, 2022). In another study, both natural and TTS speech were equally intelligible (Bione & Cardoso, 2020: 181). In terms of retention, Davis, Vincent, and Park (2019) found no significant difference between human and machine voices (Davis et al., 2019: 9). As the study compared different prosodic variations of the human voice, it showed that decisions for the voice of an agent are more complex than human or machine (Davis et al., 2019: 10), as a weak human prosodic voice was rated better than a machine voice in comparison to a strong prosodic voice. Thus, linguistic considerations like the features to be included in the synthesis of the voice are essential.

The influence of linguistic cues and features on learning has been examined on different scales from small-scale perceptual differences in sounds to large-scale differences in rhetorical structure and pragmatic dimensions like politeness (e.g. L. Lin et al., 2020; Ylinen et al., 2010). Social information regarding dialect, age, and gender can influence the perception of sounds (review in Drager, 2010) and accordingly affect speech perception. According to Kleinschmidt, Weatherholtz, and Florian Jaeger

(2018), the key to robust speech perception in a world where there is a lack of invariance is to recognize familiar information, generalize to similar contexts and adapt to novel contexts. Thus, to ease speech perception, it may be useful to provide more familiar information and support adaptation. Cristia et al. (2012) point out that accented speech initially perturbs word recognition and sentence processing, yet with exposure, people can adapt to it (Cristia et al., 2012: 4). Thus, language speakers can adapt to any accent with sufficient exposure, yet it may be helpful to ease this adaptation.

The psycholinguistic insights on perceptual learning are interesting for us on two levels - the human vs computer dimension and the language variety dimension. Looking at the human vs computer dimension, in McAuliffe, Babel, and Vaughn (2016), perceptual learning of fricatives by native English speakers took place regardless of whether participants were exposed to resynthesized or naturally produced speech (McAuliffe et al., 2016: 4). It seems that synthesized language does not hinder perception. Still, it should be noted that this is a perceptual learning task applied to sounds and thereby not fully comparable with our learning scenario aiming to test the retention and transfer of knowledge. Regarding the language variety dimension, it has been demonstrated that dialect information is extracted pre-attentively and is divided in- and out-group categories (Scharinger, Monahan, & Idsardi, 2011). If the cues of the PA indicate a particular variety, this is likely to be rapidly perceived and categorized as such. It is also expected that the variety will be associated with attitudes towards it.

Language attitudes are built in a complex manner. Speakers can express more positive attitudes towards in-group than out-group varieties, for example in the attitudes towards teachers speaking the same US variety as the learners in Gill (1994); yet speakers can also prefer out-group varieties, for example when non-native speakers have more positive attitudes towards native speakers. In a survey of Czech English students by Brabcová and Skarnitzl (2018), most students wanted to acquire a native English accent like British English, though they believed that students should be exposed to different accents in English lessons. In Aiello (2016)'s survey of Italian students, the statement "I don't want an Italian accent when I speak English." had an average score of 3.06 Aiello (2016: 295), which corresponds to "Slightly agree". Thus, the in-group membership in this case was not very desirable.

Attitudes are not restricted to general preference, but they can influence what other characteristics are ascribed to speakers of the variety such as open-mindedness and aggressiveness but also content credibility, i.e. whether information delivered by a speaker of the variety can be trusted. For example, in a study evaluating a tour guide speaking Standard Austrian or Viennese Austrian, the Standard Austrian guide was classified as educated, trustworthy, competent, polite, and serious, whereas the Viennese Austrian guide was rated most natural, emotional, relaxed, open-minded, with the highest sense of humor, but also as most aggressive (Krenn, Schreitter, & Neubarth, 2017: 75). Voice assistants are consistently attributed age and race characteristics based on their vocal cues as well as personality traits (funny, (in)competent), which correlate with common stereotypes (Holliday, 2023).

Dahlbäck and colleagues2001; 2007 found evidence for the similarity-attraction effect in the perception of information read in Swedish and American English, as listeners trusted information spoken from a speaker of their own variety regardless of whether it was trustworthy due to signs of higher expertise or not. Evans and Michael (2014) showed that true and false statements produced by non-native Hispanic speakers of English were judged by (non-Hispanic) raters to be more deceptive than those produced by native speakers.

Lev-Ari and Keysar's well-replicated 2010 study found that trivia statements produced by speakers of non-native varieties are less trusted than those produced by speakers of native varieties. Czech students also rated the native speakers more credible than the Czech non-native speakers of English

(Hanzlíková & Skarnitzl, 2017; Podlipský, Šimáčková, & Petráž, 2016). However, Souza and Markman (2013) were not able to show that foreign accent is less trusted. A later study replicating Lev-Ari and Keysar (2010) proved that with increased exposure, people who have had no contact with a non-native variety like Polish English can trust it even more than a native variety like British English (Boduch-Grabka & Lev-Ari, 2021). Moreover, Italians were found to trust statements pronounced by non-native Italian speakers more than those uttered by native Italian speakers (Lorenzoni, Pagliarini, Vespignani, & Navarrete, 2022). Testing whether Spanish speakers are likely to respect new social norms such as COVID regulations spoken with a native Spanish vs a non-native Arabic accent, Bazzi, Brouwer, Planelles Almeida, and Foucart (2022) found that the likelihood of people respecting new norms did not depend on the accent (Bazzi et al., 2022: 15). Thus, the incorporation of non-native varieties in PAs may have promising outcomes.

6.2.3 Credibility and intentionality

Credibility has been viewed to contain several dimensions - "competence", "trustworthiness" and "goodwill" ("intent toward receiver") (McCroskey & Teven, 1999: 90). The good intent when someone is credible evokes another concept - intentionality. According to Prinz (2017), intentionality indicates that "conscious experience always implies the experience of a particular kind of access to a particular kind of content" (Prinz, 2017: 348), for example, hearing implies hearing something, such as hearing a sound. In the case of designing PAs, intentionality motivates the conscious integration of features that aim to evoke something, e.g. an impression of a credible agent. The use of a non-native variety in our study aims to give the PA some intentionality - it invokes pre-formed intentional representations in users' minds. As users perceive the PA's intent to improve their experience, they are more likely to attribute to it positive characteristics, to trust it, and to want to work with it. Many non-native English speakers from the expanding circle have learned English in a formal education environment such as a local school, so they are likely to have had the most contact with their classmates and teachers, who are also likely to be speakers of their local English variety. Thus, the local variety is their in-group variety (for in- vs out-group, see review in Molenberghs, 2013) and the more familiar variety, so when it is incorporated in a PA learning environment, this may have a social mirroring effect (Prinz, 2013). The agent mirrors its user and if they share a conceptual framework, this act is perceived by the user (Prinz, 2013: 1107). The user attributes intentionality to the PA even though it does not have the prerequisites (i.e. a brain equivalent to the human brain) to develop intentionality (cf. Searle, 1980). As a result, the user understands the intentionality of the agent in its attempt to be like the user and be liked by the user, which can lead to a positive perception on the side of the user.

Linguistic credibility becomes an expression of goodwill, a form of caring. As previous research has shown, the (human) teacher's care for the students and their good relationship impacts the teacher's credibility (Teven, 2007: 435). So how can intended goodwill be attributed to a machine? There are of course many factors to consider, but from our linguistic perspective, it is important for the agent to mirror the sociolinguistic environment of the learners. The studies described above have repeatedly demonstrated the positive effect of familiarity. In the context of PAs, Maes (1994) has also argued that the agent should adapt to the user needs by imitating the user and implementing their feedback (Maes, 1994: 40).

Intentional design means taking care of individual differences of the users. For example, for groups preferring directness, the text should not be made unnecessarily polite, as this could lead to negative face

threat (P. Brown & Levinson, 1987). Also, when a non-native variety is mirrored, it is important to avoid face threat through stereotypes (Labov, 1994: 78). The TTS voice should be a credible representation of the variety and not a caricature. Finally, integrating non-native varieties in educational products has the potential to improve the attitudes towards them and increase their acceptance, leading to diversification and challenging negative stereotypes (Sutton et al., 2019: 10-11). Attributing non-native varieties credibility is again a form of goodwill and caring.

Overall, creating a credible agent that gives the impression of intentional goodwill for the user means making the agent customized to the needs of the user. The design has to be suitable for individualization but should also be context-aware (Sutton et al., 2019: 6-7). In the context of language varieties, the agent should mirror the variety but also avoid evoking stereotypes. It has become clear that more attention should be paid to the influence of socio-cultural variables when designing language technologies. We attempted to apply the principles discussed in this section to the design of a TTS system speaking non-native varieties and tested it in a lecture scenario. The following sections describe the PA implementation and discuss the results of the study.

6.3 Methods

6.3.1 participants

We recruited 82 Italian advanced speakers of English at partner universities in Italy. They were compared to a pilot study with 20 English speakers from the Czech Republic. We excluded 14 participants who selected "Other" as their country of origin and analyzed the data of the remaining 74 Italians and 15 Czechs. All participants received course credit for their participation. The students were mostly between 21-23 years old (76%) and 24-26 years old (13%) and mostly females (70 females, 19 males). They mostly came from the field of Humanities (80% of the Czechs, 85% of the Italians) and were in their Master's, except for one Bachelor's and 9 PhD students in the Italian sample. Almost all students were in their first year (93% Czech, 97% Italian). The students had a similar level of English - about 15 years on average, where the Czech participants studied English for 16.9 years (SD = 4.53) and the Italian participants studied English for 15.2 years on average (SD = 2.81). In addition, only one student had spent time in an English-speaking country, thus the level of English should be similar across the samples. Since most were students of English, they are expected to have comparable background knowledge with regard to the topic of content mapping. Prior knowledge is a variable difficult to control but essential for keeping a comparable cognitive load, which would help avoid confounds and explain the learning outcomes and performance (Davis et al., 2019: 3).

Figure 6.1 shows a similar exposure to English and communication in English comparing the Czech and the Italian participants. The Italians have had more native speakers at school while the Czech use English more in online games and at work, yet this is unlikely to impact the learning test results. In terms of the use of English at university, which is the relevant factor for our learning task, the groups report similar use of English, as both use English at university and have had native English speakers at university.

The questionnaire also gathered data on the technology affinity of the participants. Figure 6.2 presents a summary of four key variables: subplot (A) demonstrates a higher use of virtual assistants in the Italian than in the Czech group. Subplot (B) also shows more variable contexts of virtual assistant use in the Italian group, yet both groups predominantly use assistants on their phones. Subplots (C)

and (D) show consistent evaluations of technology affinity, as technology proficiency (C) has been evaluated with predominant agreement whereas lack of interest in technology (D) has been evaluated with predominant disagreement.

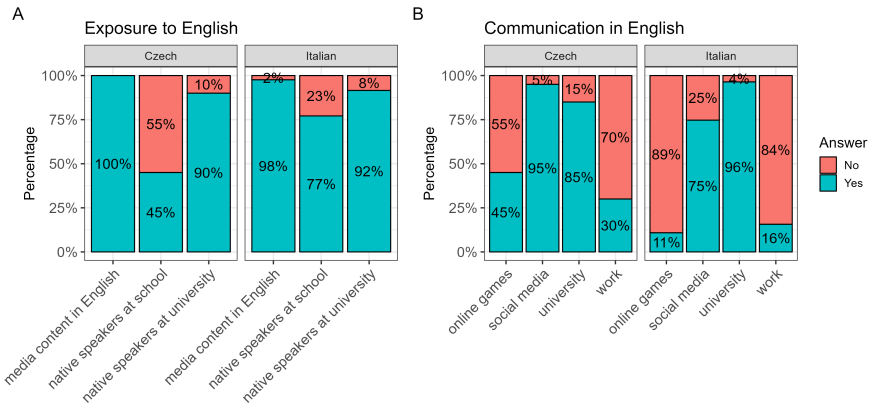


Figure 6.1: (A) Situations of exposure to English and (B) places of communication in English of the Czech and Italian participants.

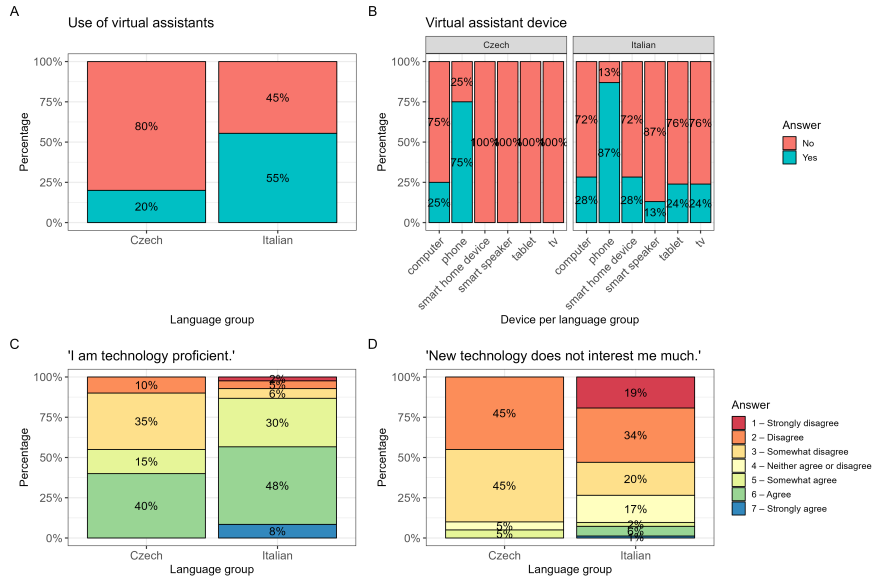


Figure 6.2: Tech-savviness as represented in: (A) Overall use of virtual assistants, (B) Devices used with virtual assistants, (C), Self-evaluated technology proficiency, (D) Self-evaluated interest in technology.

6.3.2 Materials

The lecture incorporated an excerpt from a Linguistics article by Delibegović Džanić and Berberović (2021) "Lemons and Watermelons: Visual Advertising and Conceptual Blending" (see Appendix). The excerpt explains cognitive content mapping on the basis of a joke "If life gives you lemons, a simple surgery can give you melons" (Delibegović Džanić & Berberović, 2021: 124). It is difficult to find measures for the text's complexity because most of them are designed for native English texts, and are readability measures, which may not transfer to listening. Using the Python library textstat (Bansal & Aggarwal, 2022), we compared different readability scores. However, the scores were relatively contradictory (Appendix table 6.3), as EFLAW, a score usually used for second-language learners was 24.2, which is between 20.49 and 25.49 and thereby quite easy to understand, whereas many of the native readability scores placed the text in the 11th grade, which is relatively complex. Still, despite these discrepancies in automatic complexity measurements, a manual evaluation of the text shows that the lecture is complex due to its high-tier academic vocabulary and structure and even more so when it is only heard. Thus, the lecture was able to challenge the participants and prompt learning, which can be measured through their responses on the test.

For the synthesis of the text, we used the TTS system developed in the frame of the project "Credible Conversational Pedagogical Agents" part of the Collaborative Research Center "Hybrid Societies". The text was transcribed and synthesized in Italian English with the 6208-IPA-3500 [Italian] model, Czech English with the tacotron-TZ-IPA-6000 model, and American English with the tacotron-LJS-IPA-101500 model. The American English model was trained for 101500 iterations on the LJSpeech dataset, whereas the Italian and Czech models were first trained for the same amount of iterations on the LJSpeech dataset and then for 3500 and 6000 iterations on their respective datasets. The text was first synthesized to a Mel spectrogram using tacotron (Taubert, 2022b), then the Mel spectrogram was synthesized to a WAVE audio file with waveglow (Taubert, 2022c) and finally, the audio was denoised with denoiser (Défossez, Synnaeve, & Adi, 2022). The models were trained on a corpus of annotated sociolinguistic recordings with Czech and Italian English speakers. Since the recordings were also of university students in their 20s, the PA has a relatively young-sounding voice, which may impact its perceived credibility (Beege, Schneider, Nebel, Mittangk, & Rey, 2017), but this is probably unlikely, since the lecture is not designed to evoke distrust.

The synthesized lecture lasted 2:30 min. The short length of the listening text was justified by factors such as attention span and memory capacity, as elaborated in S. D. Craig and Schroeder (2019: 1545). While there is some evidence for the benefits of simultaneously presenting audio and graphical representation (Davis et al., 2019: 3), we incorporated only audio to avoid confounding multimodal effects that would be generated by complementary text, images, or embodied speaker, and to focus only on the voice's variety effects. As Davis et al. (2019: 3) point out, an advantage of presenting only one mode is that cognitive processing channels are not overloaded by simultaneous input. In addition, Mayer's meta-analysis showed that adding the speaker's image is not very effective (Mayer, 2014b: 360).

The participants were mostly able to identify the speaker of their own variety but had issues recognizing the American English variety (figure 6.3). Note that we use the abbreviations AmE_Cz for American English evaluated by Czechs, AmE_It for American English evaluated by Italians, CzE for Czech English, and ItE for Italian English. Among the wrong responses, Italians considered the AmE speaker to come from England and the UK whereas the Czechs guessed Germany or Denmark. There were also a couple of answers indicating that this is a computer-generated voice. Generally, people use

their knowledge of linguistic and social variables' statistical distribution in the world to infer about the listened speech and classify it into social categories (Kleinschmidt et al., 2018: 829). The cues in the American TTS voice allowing it to be classified as American were either not strong enough or the participants may not have had sufficient exposure to the variety. While this may be considered a disadvantage for our TTS system, it is partially an advantage for the experiment, as we expect that the origin of the speaker has not biased the ratings of the characteristics in the AmE versions.

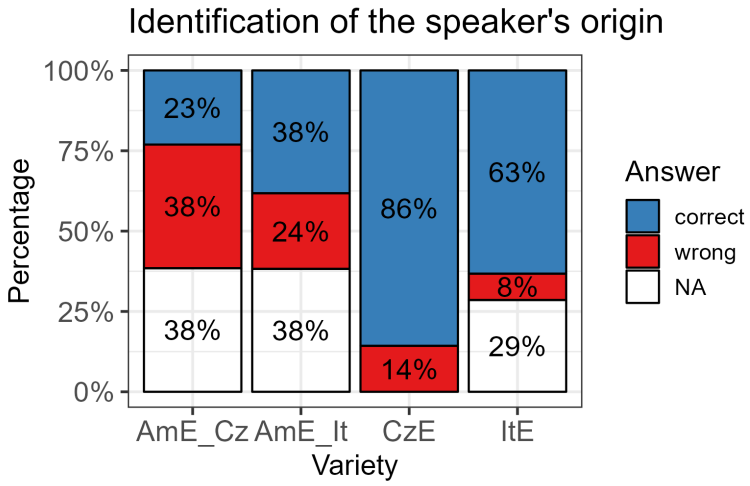


Figure 6.3: Proportion of correct and wrong answers on the speaker's origin per variety.

6.3.3 Procedure

The survey was conducted online on the LimeSurvey servers of the Chemnitz University of Technology. All questions were obligatory. The survey opened with information on the background and goals of the research and on data protection together with an informed consent form. Then the participants were randomly assigned to listen to the lecture produced either in the local variety (Italian English or Czech English respectively) or in American English. Afterward, the participants rated the following statements on a five-point scale (where five is the highest): "The speaker is credible.", "The speaker is warm.", "The speaker is competent.", "The speaker is easy to understand.", and "I like the voice of the speaker". They were also asked whether they had a feeling where the speaker could be from.

In the test component, the participants responded to five general vocabulary questions, five academic vocabulary questions, and three transfer learning questions (see Appendix). We aimed to combine different question types to test retention and transfer (see e.g. Buck, 2007). The general vocabulary items were multiple-choice questions with one correct and two wrong answers, which inquired about high-tier vocabulary from the text such as "body contouring" and "astringent". These questions served to test and compare the English level of the participants. The second set of questions, the academic vocabulary type, were again multiple choice, but aimed to test the understanding of core linguistic concepts from the lecture such as "content mapping" and "input space". The final three questions on transfer learning included one short and one long open question and one multiple-choice question, which tested the application of the knowledge acquired in the lecture. The participants had to explain jokes similar to the one used in the lecture, which prompted the application of the new concepts.

After the test section, a demographic questionnaire was presented. It covered background data like age, field of study, gender, country of origin, and foreign languages as well as questions on tech-savviness and linguistic experience (see Appendix). The demographic questions were left to the end in order to keep the primary position for the test (see Dörnyei, 2003: 61) and to avoid biasing the participants about the synthetic nature of the voice due to the questions inquiring about tech-savviness.

6.3.4 Analysis

While the multiple-choice questions could be evaluated in a simple true-false manner, the open questions were rated on a five-point scale based on the correctness of the responses and the grasp of the information provided in the lecture, where five was the highest value. This allowed us to give credit to responses, which have only partly grasped the ideas of the lecture. The responses of the test and the demographic data were analyzed in R via RStudio (R Core Team, 2023; RStudio Team, 2023) with the tidyverse (Wickham et al., 2019) and rstatix packages as well as the core stats package. Chi-square tests were applied to the categorical data of the correct and wrong responses to the multiple-choice questions. For the grades of the open questions and the rating of the voice's characteristics, after testing the assumptions for normality via Shapiro-Wilk tests and homogeneity of variances with Levene's test, the comparisons were conducted with the non-parametric Wilcoxon signed-rank test. Isolating only the Italian sample, the influence of the different lecture varieties was compared with an unpaired two-sample t-test.

6.4 Results

6.4.1 Learning outcomes

The learning outcomes using the PA were evaluated based on the participants' test results. An overview of the proportion of correct and wrong responses per lecture variety is presented in figure 6.4. The red, i.e. incorrect responses appear more pronounced in the Czech sample, both in Czech English (CzE) and AmE_Cz. A two-sample chi-square test confirmed that the Language groups differ in terms of answer correctness [$\chi^2(1, N = 979) = 18.8, p < .001$]. Looking at the standard residuals of the correct and wrong answers per language group, it becomes evident that the Italian group has performed better than the Czech group (Appendix table 6.4). There is also a significant difference between the impact of the different lecture varieties on the correct and wrong responses: [$\chi^2(3, N = 979) = 20.8, p < .001$]. Again, as expected from the test on language group and variety, there are overall more correct responses in the Italian group than in the Czech group. The standardized residuals of the correct responses with the Italian English variety (2.87) seem to be higher than the correct responses with the American English variety (0.5) (Appendix table 6.5). Yet, when only the Italian data is tested, the number of correct and wrong responses does not differ based on the variety of the lecture [$\chi^2(1, N = 814) = 0.646, p = .422$]. Since one of the aims of the general vocabulary questions was to test the language proficiency of the test takers, a chi-square test only on the responses of the general vocabulary questions determined that the groups differ in their answer correctness [$\chi^2(1, N = 445) = 10.7, p = .001$]. The standardized residuals table shows that the Italian group has performed slightly better than the Czech (Appendix table 6.6), similar to the results of the overall table on the combined multiple choice questions (Appendix table 6.4).

Looking at the results from the open questions, there is no difference between the grades of the Czech (M = 2.4, SD = 1.43) and Italian (M = 2.77, SD = 1.27) groups [$W = 473, p = .368$]. There is also

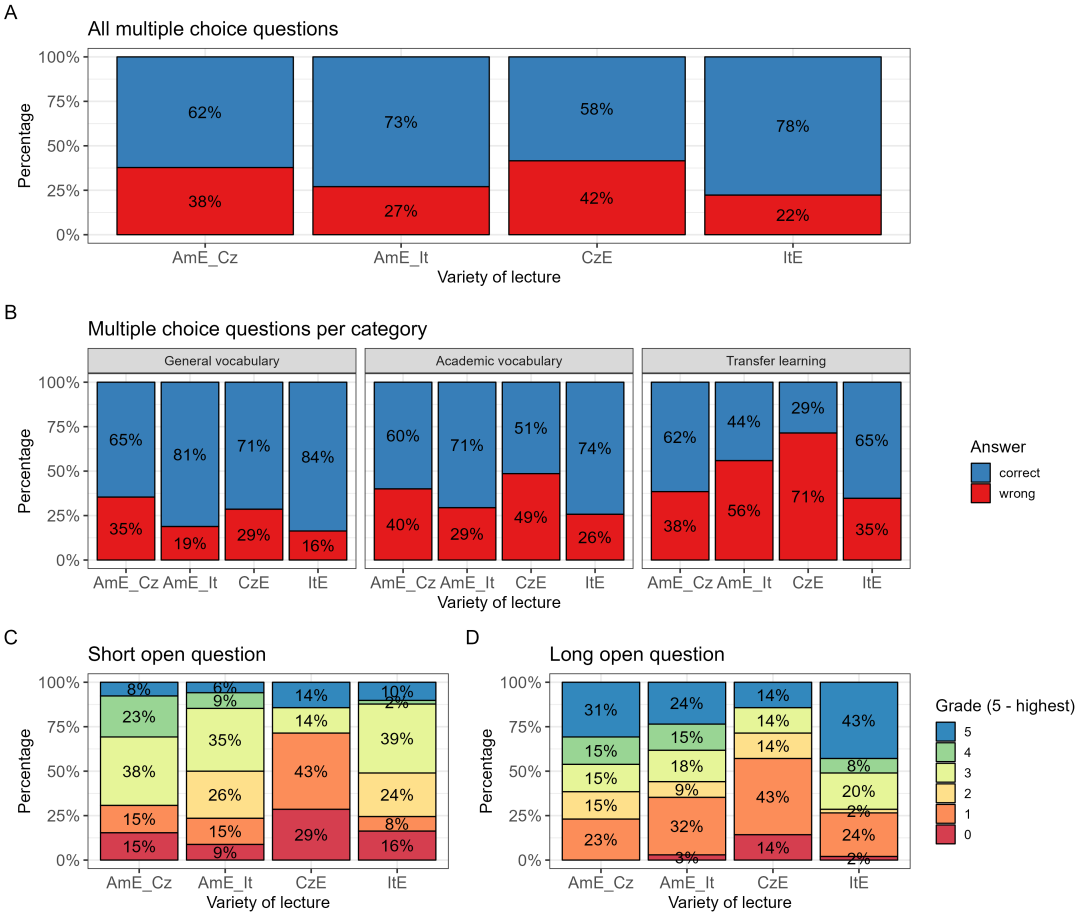


Figure 6.4: Overview of the responses to the multiple choice and open questions per lecture variety.

accordingly no difference between the scores based on the variety of the PA. When only the Italian data are concerned, there is again no difference in the open question grades under the lecture variety condition [$t(72) = -0.817, p = .416$].

We tested whether other variables influence the test results. There was no difference between the responses of the students who use virtual assistants and those who do not [$\chi^2(1, N = 979) = 0.165, p = .685$]. Study degrees also did not play a role as students from the BA, MA, and PhD levels had a similar response ratio of about 3:1 correct vs wrong responses.

6.4.2 Characteristics

The Czech and Italian participants have rated the characteristics similarly on average [$W = 452, p = .259$]. The mean rating by Czechs (M = 2.81, SD = 0.63) and Italians (M = 3.07, SD = 0.885) is close to the middle of the five-point scale. The speaker characteristics themselves have been rated differently, as pairwise Wilcoxon tests showed that the COMPETENCE characteristics *competence*, *credibility*, and *ease of understanding* are rated higher than the WARMTH characteristics *warmth* and *likableness* (table 6.1).

Group 1	Group 2	N1	N2	W	p	p adj.	Sign.
competence	credibility	89	89	3894.5	0.843	1	ns
competence	ease of understanding	89	89	3657.5	0.364	1	ns
competence	likableness	89	89	5411	1.34×10^{-5}	8.04×10^{-5}	***
competence	warmth	89	89	5760.5	5.64×10^{-8}	5.08×10^{-7}	***
credibility	ease of understanding	89	89	3691.5	0.42	1	ns
credibility	likableness	89	89	5435.5	1×10^{-5}	7×10^{-5}	***
credibility	warmth	89	89	5780.5	4.21×10^{-8}	4.21×10^{-7}	***
ease of understanding	likableness	89	89	5422.5	1.37×10^{-5}	8.04×10^{-5}	***
ease of understanding	warmth	89	89	5709.5	1.77×10^{-7}	1.42×10^{-6}	***
likableness	warmth	89	89	4159.5	0.548	1	ns

Table 6.1: Results of Wilcoxon signed-rank tests comparing characteristic ratings.

In terms of variety of the PA, The Italians have rated AmE higher relative to Italian in comparison to the Czechs' rating of AmE relative to CzE. This is evident from the results of the pairwise Wilcoxon tests in table 6.2 and the boxplots with error bars in figure 6.5. We also looked at participant gender, which was shown to affect evaluations of warmth and competence of language varieties in McKenzie et al. (2016). However, gender did not affect characteristics ratings in our study [$W = 770$, $p = .291$].

Group 1	Group 2	N1	N2	W	p	p adj.	Sign.
AmE_Cz	AmE_It	8	27	59	0.056	0.223	ns
AmE_Cz	CzE	8	7	31	0.768	1	ns
AmE_Cz	ItE	8	47	208.5	0.631	1	ns
AmE_It	CzE	27	7	155	0.01	0.052	ns
AmE_It	ItE	27	47	992.5	5.61×10^{-5}	0.000 337	***
CzE	ItE	7	47	150.5	0.726	1	ns

Table 6.2: Results of Wilcoxon signed-rank tests on characteristic ratings per variety.

6.5 Discussion

This study examined the effect of the use of a PA in teaching a lecture on content mapping in linguistics. Although we expected to find improved learning outcomes after a lecture spoken in the non-native variety, we were not able to find evidence for the effectiveness of a PA speaking English varieties. There might be a narrow window of learning material complexity (neither too easy, nor too difficult) for PAs credibility to influence learning outcomes. Our findings are similar to Mayer, Dow, and Mayer (2003), where student performance under an American-accented and Russian-accented voice did not differ significantly in terms of retention (Mayer, Dow, & Mayer, 2003: 422). One reason may lie in the complexity of the lecture, as it may have been more suitable to be read than spoken. In future studies, the agent could be adapted to various oral presentation styles like a podcast. Speaking style is decisive for the perception of instructor credibility with many different influential variables like speaking

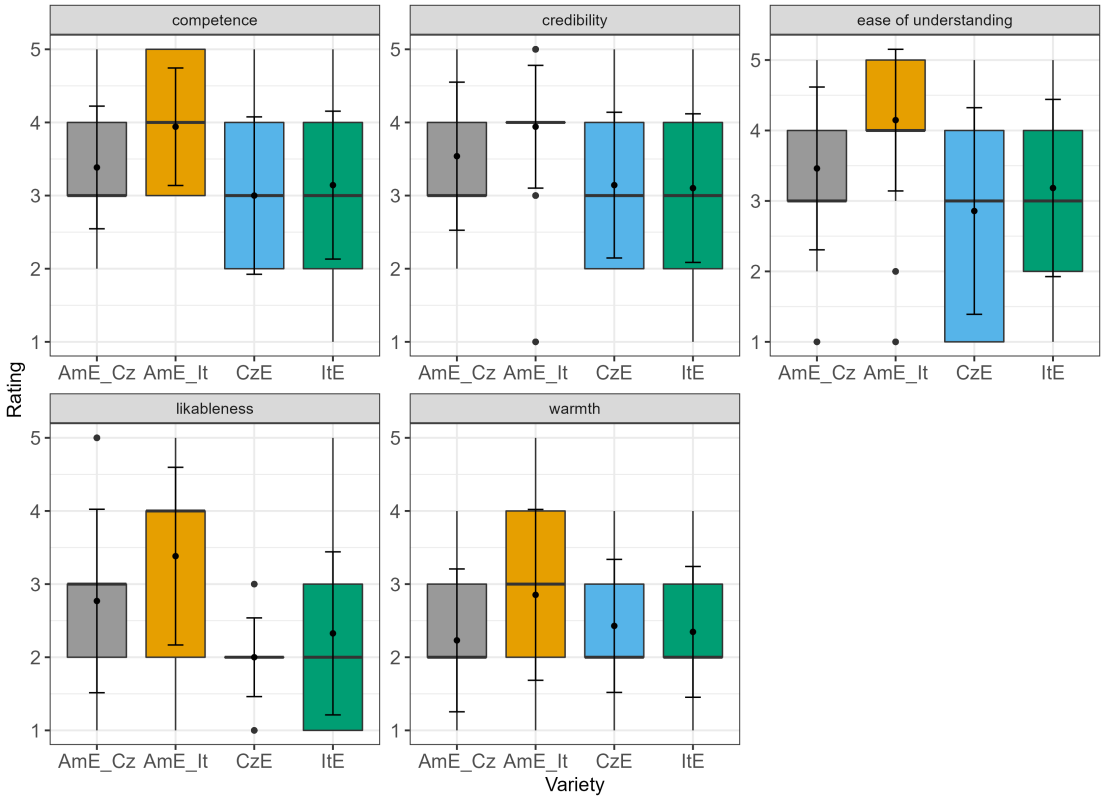


Figure 6.5: Voice characteristics rated by variety of PA. Error bars represent standard error.

rate and the use of humor (see the review in Myers & Martin, 2018). As underscored in the review by S. Schneider, Beege, Nebel, and Rey (2022), emotional factors like enthusiasm expressed through voice cues like intonation are also likely to benefit the learning process, yet this depends on the mental load during the learning task (S. Schneider et al., 2022). There are thus many adjustable parameters in the lecture presentation that could affect the learning outcomes. The lecture format as a whole can also be modified, as instructor credibility can be influenced by structural classroom components such as course format (Myers & Martin, 2018: 41). In the process of lecture improvement, the biggest challenge would be to reduce confounds. Schroeder and Gotch (2015: 189) demonstrate that one of the biggest methodological issues of studies using PAs is that they fail to isolate the effects of the PA, where confounds affect the results on the level of the experiment treatment and control conditions and the design of the PA. Thus, if the lecture is made too conversational, this may switch the focus from testing the voice principle to testing the personalization principle.

Another limitation of this study is that while the incorporation of only one mode (audio) was useful for reducing confounds, it reduced the ecological validity of the experiment, as learning with an audio-only narration is far from a real-life lecture scenario. Students can usually assist their comprehension through note-taking and visuals like presentation handouts. However, while the addition of text seems intuitive, it could contradict the redundancy principle and become distracting. Moreover, even though the lecture was short, it proved difficult to follow. This complexity was necessary to test the transfer

learning effects on novel input and isolate background knowledge confounds. Managing complexity remains one of the core challenges of educational design and it opens promising opportunities for follow-up research.

In terms of the groups' learning outcomes, the Italian participants performed better than the Czech in the multiple-choice questions, but the groups scored similarly in the open-choice questions. The isolated analysis of the general vocabulary questions confirmed that there may be a hidden English level difference of the language groups, which is not reflected in their otherwise similar number of years studying English. One of the main limitations of this study was the different sample sizes of the two groups, therefore, it would be useful to improve the sample by testing more Czech participants.

While the Italians reported using virtual assistants more than the Czechs, this did not impact their learning outcomes in the survey. Since all students evaluated themselves similarly in terms of technology affinity, it is understandable that this did not affect their results. Future studies can either incorporate participants with larger differences in technology affinity or exclude this question to make the survey shorter and help reduce potential neutral response bias.

When it comes to the evaluation of the speaker's voice, the Czech and Italian participants rated the voice characteristics similarly on average - around the center of the rating scale. This may be a result of neutral response bias perhaps with some influences of social desirability bias. Still, we found a difference in the evaluation of the characteristic clusters: the COMPETENCE characteristics (*credibility*, *competence*, *ease of understanding*) were rated higher than the WARMTH characteristics (*likableness*, *warmth*). These ratings did not differ in terms of rater gender, contrary to the findings in McKenzie et al. (2016) on Thai attitudes to English varieties, possibly due to differences in the sample as well as differences between their human voice and our synthetic voice. Our computer voice appears to be consistently perceived as competent but rather unlikable, thus it would be necessary to improve its prosody and thereby its naturalness Ehret et al. (e.g. 2021). S. D. Craig and Schroeder have twice found that modern computer voices do not differ in terms of ascribed credibility compared to human voices, yet human voices are rated higher in terms of human-likeness and engagement (S. D. Craig & Schroeder, 2017, 2019: 1542). Schroeder and colleagues have also shown that the more human-like the voice, the more consistent it is rated in terms of trust responses (Schroeder, Chiou, & Craig, 2021: 11). Still, trust in the virtual speaker had small influence on learning (Schroeder et al., 2021: 12). These findings may be in a way reflected in our results as well, as competence variables like credibility were rated higher than warmth variables such as likableness. Similarly, the relatively high credibility ratings also did not influence learning outcomes. Since our study compared two voices of the same agent, we cannot determine whether the reason for the ratings and learning outcomes lies in the voice quality of the agent. It would be interesting to compare the PA to other agents and to a human voice. Also, while the perceived young age of the speaker most likely did not affect its credibility perception, it would also be important to control for the age variable (see Beege et al., 2017).

The higher responses in favor of credibility raise the question of whether the ratings have been influenced by the academic language of the lecture. One solution would be to test students from a lower study degree with an easier text. The underrepresentation of K-12 students in PA research has been highlighted (Schroeder, Adesope, and Gilbert 2013: 20; Schroeder and Gotch 2015: 192), as there is some evidence that they profit from PAs more than post-secondary students (Schroeder et al., 2013). The computer voice in other studies was also rated less dynamic and attractive (Mayer, Sobko, & Mautone, 2003: 423), so the overall lower rating of likeableness may be reflecting attitudes towards the computer voice.

Our findings on the preference for American English by the Italians are similar to Mayer, Sobko, and Mautone (2003), where the American English speaker was also rated more positively than the Russian-accented speaker. Italians have rated AmE higher than Czechs when comparing it to Italian English (ItE) and CzE respectively. The Italians may have more negative attitudes to ItE than the Czech to CzE. As discussed in Reichelt et al. (2014: 207), the positive effect of personalization can be modulated by the content, learning phase duration, cognitive load, and some individual variables like previous knowledge. In our case, the attitude towards English varieties has probably affected the users' evaluation of the voice but it is not necessarily obvious whether it has influenced the learning outcomes. If the preference towards AmE were beneficial, we would have observed better learning outcomes in that variety, yet the test results were not found to differ based on variety.

While in our study cognitive load was modulated implicitly by increasing the difficulty of the tasks, psychophysiological methods like electrocardiography (ECG) or a combination of methods like ECG and electroencephalography (EEG) (Jimenez-Molina, Retamal, & Lira, 2018) can provide additional insights on the cognitive load during different language varieties and test tasks. For instance, Baceviciute, Lucas, Terkildsen, and Makransky (2022) have combined EEG and eye tracking to show that in virtual reality, redundancy does not make input more cognitively demanding. Incorporating such methods can help measure cognitive load more reliably and determine the effectiveness of the applied educational design principles.

6.6 Conclusion

Technology is situated in a sociolinguistic environment and needs to be adapted to it. Every speaker has an accent, so designers of speech technologies always face the choice of which model they want to incorporate. While so far the explored varieties have predominantly been native English varieties like American English, we tested whether a PA speaking the user's own non-native variety can improve learning outcomes in an English-medium environment, in accordance with the voice, personalization, similarity-attractiveness, and intentionality principles. Experiments with Italian and Czech groups showed that the variety of the agent did not significantly improve the learning outcomes. Still, we found different evaluations of characteristics related to the warmth and competence of the PA where the speaker was rated more credible than warm or likable. Italians rated AmE in relation to Italian English higher than the Czechs, indicating more positive attitudes towards AmE. With improved TTS model prosody and a larger participant sample, future studies can test the influence of variety familiarity on learning. By incorporating intentionally credible voices, language technology developers can address the individual needs of learners from different backgrounds and thereby give them active control over the product and the learning process. These efforts have the potential to ease the learning of non-native English speakers in English-medium instruction learning environments by adapting to their individual varieties of English.

6.7 Acknowledgements

We thank Olga Dontcheva-Navratilova for the dissemination of the surveys at Masaryk University Brno and the discussion of the results. We are also grateful for all students who participated in the survey. We thank the team of the D03 subproject of the Collaborative Research Center "Hybrid Societies" at Chemnitz University of Technology for their contributions in the development of the

TTS system and their inputs on the learning experiment. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID4 16228727 - SFB 1410, TP D03. This work was also supported by a PhD scholarship awarded to MI by the State of Saxony.

6.8 Appendix

6.8.1 Supplementary tables

Readability measure	Value	Interpretation
EFLAW	24.2	Quite easy to understand
Flesch	61.87	Standard
Flesch-Kincaid	9.1	Grade
Fog	11.61	Grade
SMOG	12.7	Grade
Automated Readability Index	12	11th grade
Coleman-Liau Index	11.77	Grade
Dale-Chall Readability Score	8.83	11th or 12th-grade
Difficult words	79	
Consensus		11th and 12th grade

Table 6.3: Comparison of readability measurements of the presented lecture's text.

Language Group	Test answered	Observerd	Proportion	Row proportion	Column proportion	Expected	Residuals	Std. residuals
Czech	correct	100	0.10	0.61	0.14	122.70	−2.05	−4.44
Italian	correct	628	0.64	0.77	0.86	605.30	0.92	4.44
Czech	wrong	65	0.07	0.39	0.26	42.30	3.49	4.44
Italian	wrong	186	0.07	0.23	0.74	208.70	−1.57	−4.44

Table 6.4: Results of the chi squared test on the answers of the multiple choice tests per language group.

Language Group	Test answered	Observerd	Proportion	Row proportion	Column proportion	Expected	Residuals	Std. residuals
AmE_Cz	correct	55	0.06	0.63	0.08	65.44	−1.29	−2.67
AmE_It	correct	224	0.23	0.75	0.31	220.85	0.21	0.50
CzE	correct	45	0.05	0.58	0.06	57.26	−1.62	−3.33
ItE	correct	404	0.41	0.78	0.55	384.45	1	2.87
AmE_Cz	wrong	33	0.03	0.38	0.13	22.56	2.20	2.67
AmE_It	wrong	73	0.07	0.25	0.29	76.15	−0.36	−0.50
CzE	wrong	32	0.03	0.42	0.13	19.74	2.76	3.33
ItE	wrong	113	0.12	0.22	0.45	132.55	−1.70	−2.87

Table 6.5: Results of the chi square test for multiple-choice test results per variety.

Language Group	Test answered	Observerd	Proportion	Row proportion	Column proportion	Expected	Residuals	Std. residuals
Czech	correct	50	0.11	0.67	0.14	60.67	−1.37	−3.44
Italian	correct	310	0.70	0.84	0.86	299.33	0.62	3.44
Czech	wrong	25	0.06	0.33	0.29	14.33	2.82	3.44
Italian	wrong	60	0.13	0.16	0.71	70.67	−1.27	−3.44

Table 6.6: Results of the chi squared test on the answers of the general vocabulary test per language group.

Group 1	Group 2	N1	N2	W	p	p adj.	Sign.
competence	credibility	74	74	2744	0.982	1	ns
competence	ease of understanding	74	74	2513	0.374	1	ns
competence	likableness	74	74	3691	0.000 17	0.001	**
competence	warmth	74	74	3971.5	9.64×10^{-7}	9.64×10^{-6}	***
credibility	ease of understanding	74	74	2490	0.327	1	ns
credibility	likableness	74	74	3652.5	3.14×10^{-4}	0.002	**
credibility	warmth	74	74	3924.5	2.51×10^{-6}	2.01×10^{-5}	***
ease of understanding	likableness	74	74	3721.5	1.16×10^{-4}	8.12×10^{-4}	***
ease of understanding	warmth	74	74	3953	1.73×10^{-6}	1.56×10^{-5}	***
likableness	warmth	74	74	2873.5	0.591	1	ns

Table 6.7: Results of Wilcoxon signed-rank test comparing characteristic ratings for the Italian data.

6.8.2 Survey

Transcript of the lecture for the learning task:

> The first image that we analyse in this chapter is an advertisement for a cosmetic surgery clinic in which typography works well with other elements of the composition. It is a computer-edited photo with slices of lemons and watermelons in the background. The central textual part of the advertisement says, "If life gives you lemons, a simple surgery can give you melons". Words lemons and watermelons are highlighted as they are printed in yellow and red ink on the black background.
>
> The central element of the advertisement is the textual part, which contains a proverb in a modified form, while the visual part, which contains various fruits, is placed in the background. The meaning of this advertisement is created in a five-space integration network. Input space one contains the canonical phrase if life gives you lemons, make lemonade. The proverb is used as a piece of advice to tell someone to focus on the good in an unfortunate situation and act accordingly. Input space two contains knowledge about cosmetic surgery, a surgical specialty dedicated to enhancing a patient's appearance. The scope of cosmetic surgery procedures includes facial and body contouring, breast enhancement and others. Input space three includes knowledge about different fruits, especially their various shapes and sizes. Cross-space mappings connect lemons from input space one and fruits from input space three, as well as breast surgery from input space two and shapes and sizes of different fruits from input space three.
>
> The blend receives projections from all three input spaces. Projected to the blend from input space one is the first part of the original proverb. Thus, the first part of the canonical form is retained, if life gives you lemons, which implies that the phrase starts with the negative force. However, it suddenly changes in the second part of the utterance in the blend by projections from input spaces two and three. Projected to the blend from input space two is breast augmentation through a surgical procedure. Projected to the blend from input space three is melon. Instead of fruit with distinctive astringent flavour, melon with a delightful,

sweet taste is offered. However, the melons are significantly bigger in comparison to lemons. The shape of both of them is often used to allude to female breasts. Therefore, inputs two and three are connected by a vital conceptual relation of representation.

(Delibegović Džanić & Berberović, 2021: 124)
Survey:

General vocabulary questions

1. *Body contouring* means
    a. body painting
    b. body form change
    c. body mapping
2. *Augmentation* means
    a. enhancement
    b. transfer
    c. reduction
3. *Astringent* means
    a. strong
    b. sweet
    c. sour
4. *Allude to* means
    a. refer to
    b. attract
    c. portray
5. *Canonical phrase* is a kind of
    a. legal quote
    b. academic citation
    c. proverb

Academic vocabulary questions

6. *Cross-space mapping* means
    a. mapping between elements in cognition
    b. translation of elements between first and second language in cognition
    c. misunderstanding
7. *Negative force* means
    a. negation
    b. negative connotation
    c. contradiction
8. *Integration network* is the cognitive web where
    a. incoherent elements are contrasted
    b. multiple knowledge fields are combined
    c. idiomatic phrases are stored

9. *Input space* is
    a. the cognitive space where metaphoric and literal meaning is blended
    b. the cognitive space where we process auditory and visual input
    c. a knowledge domain providing one layer of information to linguistic processing
10. *Typography* is
    a. the spelling system of a language
    b. the style of printed matter
    c. the study of types

Transfer learning questions

11. Content mapping can be used in advertising to create _____
12. Which features of lemons are transferred?
    a. taste
    b. smell
    c. size
13. Explain how the jokes in the following advertisements work: _____

The demographic questionnaire was adapted to the Italian and the Czech survey, e.g. with "Czech Republic" and "Czech" and accordingly "Italy" and "Italian" as home country and mother language. For the sake of brevity, we present only the Czech survey here:

Necessary data

Please choose your age group.

- 18-20
- 21-23
- 24-26
- 27-29
- older
- younger

Please select your field of study. (This is the category that your major falls into. For example, if you study English, select humanities or if you study Physics, select MINT.)

- Humanities
- MINT
- Engineering
- Medicine
- Law
- Business
- Other

Please select the degree you are studying for.

- BA
- MA

- PhD
- Other

Please select which year of your degree program you're in.

- first year
- second year
- third year
- fourth year
- graduated
- Other: _____

Please select your gender.

- Female
- Male

Please tell us from which country you are.

- Czech Republic
- Other: _____

Have you lived in another country than the one you listed in the previous question for a longer period of time? (A longer period of time means for more than a year.)

- Yes
- No

Please list the countries other than the one the one you are from that you have lived in for a longer period of time.: _____

What language is your first language?

- Czech
- Other: _____

Do you speak any other languages or dialects? (This can be other local languages or dialects or other foreign languages (e.g. French).)

- Yes
- No

What other languages do you speak? (This can be other local languages or dialects or other foreign languages (e.g. French).) _____

Tech-savvyness

Do you use virtual assistants? (For example: Google Home, Apple Siri, Amazon Alexa)

- Yes
- No

On which devices do you use them? (Check all that apply)

- phone
- tablet
- computer
- TV
- smart speaker

- smart home device
- Other: _____

How often do you use virtual assistants? (Check all that apply)

- several times a day
- several times a week
- several times a month
- Other: _____

Please select how appropriate the following statements describe you.

- I am technology proficient.
  1 - Strongly disagree
  2 - Disagree
  3 - Somewhat disagree
  4 - Neither agree or disagree
  5 - Somewhat agree
- New technology does not interest me much
  1 - Strongly disagree
  2 - Disagree
  3 - Somewhat disagree
  4 - Neither agree or disagree
  5 - Somewhat agree

Linguistic experience

For how many years have you been learning English? _____

Did you have classes with native speakers at school?

- Yes
- No

Do you have classes with native speakers at university?

- Yes
- No

Do you consume media content in English? (This includes movies, TV shows, music and video games.)

- Yes
- No

Where do you interact with people from other countries? (Check all that apply)

- at university
- at work
- social media
- online games
- Other: _____

7 Conclusion

The following section summarizes the results of the studies presented in chapters 2 to 6. Afterward, the contributions of this thesis to filling the research gap identified in section 1.5 are discussed. Then, general limitations are discussed, in particular relating to underlying theoretical assumptions and methodological considerations. Lastly, possible directions for future research are presented.

7.1 Summary of Results

The systematic review of previous work presented in chapter 2 resulted in a proposed categorization of phonological features of Chinese English according to the eWAVE categorization scheme and in a proposed pronunciation model similar to the Pronunciation Model for Hong Kong English published in the online edition of the Oxford English Dictionary. While the work on phonetic features of ChE is highlighted here in line with the focus of this thesis, the review also summarized previous work on lexical and grammatical features of Chinese English[1].

The vowel space analysis presented in chapter 3 can most illustratively be summarized by the plot in figure 7.1. The analysis has confirmed previous claims that the Chinese English tense and lax vowel areas are rather similar in size, more so than most native varieties. Furthermore, the CLOTH vowel is relatively far back, which has only been reported in one recent previous study (Q. Wang, 2023). The analysis also showed that the general separability of vowels is rather low, as indicated by the large error ellipses shown in figure 7.1, that are consistently large in all subsets of the data, which is in line with previous work and indicates a general lack of the tense-lax distinction.

The work presented in chapter 3 could quantify several features of Chinese English synthesized from previous research in chapter 2. As table 7.1 shows, previous work suggested that the Chinese English vowel inventory only contains the full vowels. This is confirmed by the tentative Chinese English vowel inventory, shown in figure 7.2 and the summary of the merger analyses in table 7.2 both of which are reprinted here from chapter 3.

Keyword	Symbol	Confidence
KIT	i	high
FOOT	u	low
FLEECE	i	high
GOOSE	u	medium

Table 7.1: Excerpt of Table 1 from chapter 2, showing the categorization of phonological features of Chinese English in analogy to the eWAVE categorization

The results show that in Chinese English spoken by the participants from Guangdong, Guangxi, and

[1] It seems that the article stuck a chord with the academic community, as it was one of the top downloaded articles in *World Englishes* in 2021.

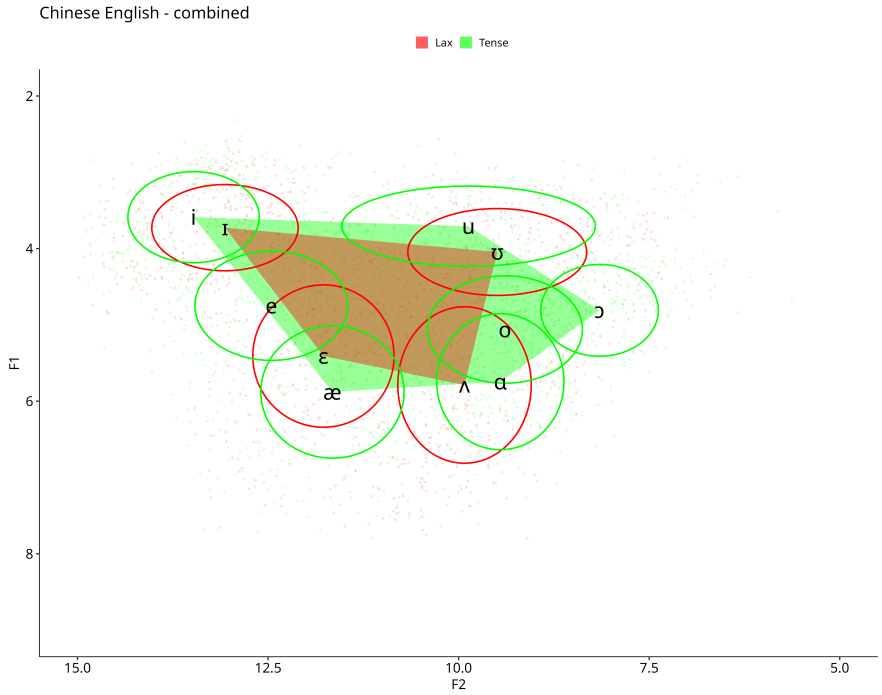


Figure 7.1: Chinese English vowel space with tense and lax vowel areas for the combined data set

Yunnan, both the FLEECE and KIT vowels and the GOOSE and FOOT vowels merge. This statement can be made with relatively high confidence since the merger analyses have revealed the mergers in the combined Chinese English data set, as well as in several subsets of the data. Yet, as table 7.2 also shows, and as confirmed by the Kruskal-Wallis test reported in chapter 3, gender seems to influence these mergers to some extent, i.e the FLEECE-KIT merger is more frequently found with male participants, while the GOOSE-FOOT merger is more frequently found with female participants.

	TRAP-DRESS		FLEECE-KIT		GOOSE-FOOT	
	Female	Male	Female	Male	Female	Male
ChE	Y/N		Y/Y		Y/Y	
ChE	Y/N	Y/N	N/Y	Y/Y	Y/Y	Y/N
Guangdong	N/N	N/N	N/Y	Y/Y	Y/Y	Y/N
Guangxi	N/N	Y/N	N/N	Y/Y	Y/N	Y/N
Yunnan	N/N	N/N	N/Y	N/Y	Y/Y	Y/Y

Table 7.2: Combined indicators of mergers (overlap analysis / SVM analysis)

The research on durational variation presented in chapter 4 quantitatively confirmed earlier impressionistic work that Chinese English exhibits different durational variation than native varieties. Furthermore, this study is the first to report a quantitative analysis of the durational variation of Czech English and

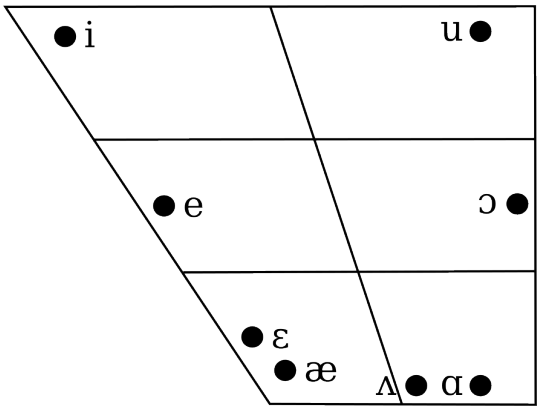


Figure 7.2: Proposed, tentative Chinese English vowel inventory

Italian English in comparison with other varieties[2]. The analysis showed that Czech English and Italian English have different durational variations, suggesting L1 influence. An alternative explanation could be different language models, as ItE is closer to AmE than CzE.

Methodologically, it has been shown in chapter 4 that most durational variability metrics produce consistent results and, while suggested in the literature, do not correlate with articulation rate or data elicitation method in our data. This allows DVMs to be used with speakers of different proficiency levels and more flexibility in eliciting data. The work presented in chapter 5 has shown that vowel space analysis can be used to construct a linguistically motivated quality metric for TTS systems[3]. This quality can be used to assess the quality of synthesized speech and inform further optimization of neural network training. In general, automated vowel (formant) analysis has been proven to be a valid method, given adequate data quality. In combination with the forced alignment of text and audio, this allows researchers to analyze large amounts of data with relative ease.

In terms of the overarching research project on *Hybrid Societies*, the results of the study presented in chapter 6 did not confirm the hypothesis that a conversational pedagogical agent speaking the same non-native variety as its users improves learning outcomes. In fact, the analysis did not show any significant difference between the American English and the Czech/Italian English condition. The part of the study that investigated participants' attitudes towards the synthetic voices revealed that the non-native voice was consistently rated as warmer and more likable than the American English voice. This finding can inform the design of embodied digital technologies to increase their acceptance.

[2] Only one previous study reported values of speech rhythm metrics (Volín & Skarnitzl, 2010), however, they investigated the correlation between these values and perceived foreign accentedness and did not explicitly investigate durational variation comparatively.

[3] The idea has been well received at the IEEE CIVEMSA 2022 conference where the paper was presented, as it was awarded the second place at the best paper awards.

7.2 Filling the Research Space

As outlined in section 1.5, this thesis tackles both the overarching research questions of the project *Credibility Through Non-native Language Varieties in Conversational Pedagogical Agents* of whether a CPA speaking non-native English is perceived positively and even enhances learning outcomes of students speaking the same non-native variety by increasing the credibility attributed to the agent. In addition, the work presented here focused on features of non-native varieties of English, in particular on the Chinese English vowel space and speech rhythm and Czech and Italian English speech rhythm. The goal was to contribute to the comprehensive, quantitative description and analysis of these varieties.

To this end, it was first necessary to systematically review existing work on features of Chinese English (see chapter 2). Based on the review, phonetic features have been classified using the occurrence frequency coding used in eWAVE (Kortmann et al., 2020). Furthermore, a list of likely grammatical features that also occur in related varieties, namely Hong Kong English (Wong, 2020) and Singapore English (Lim & Ansaldo, 2020) was compiled and thus contributes to filling the research gap.

Secondly, based on data from participants from Guangdong, Guangxi, and Yunnan, the Chinese English vowel space was investigated (see chapter 3), with the specific aim of confirming the theorized absence of vowel reduction reported in the literature. While there has been previous work on Chinese English vowels, especially from speakers from Guangxi (e.g. Deterding, 2016) and Yunnan (e.g. Ao & Low, 2016), the samples studied there were presumably convenience samples and not stratified by gender (Ao and Low 2016 did not have any male participants). No analysis of the vowels produced by speakers from Guangdong has been reported previously. The analysis showed that the data elicitation method, whether interviews or reading passages were used, only had a small effect on the vowel space. Therefore, the work presented in chapter 3 extends the work on Chinese English both in content and methodology.

Thirdly, the same data was analyzed to investigate the durational variation of Chinese English in comparison to other L1 and L2 English varieties (see chapter 4). In contrast to previous work, seven different speech rhythm metrics were used to ensure optimal comparability with data reported in the literature. The hypothesis that Chinese English exhibits less durational contrast than inner circle varieties was confirmed and it was shown that this difference is greater than the difference between English spoken by participants from different Chinese provinces. The hypothesized influence of articulation rate and data elicitation method from the literature was not found to be statistically significant. In addition, the analysis was extended to include other inner, outer, and expanding circle varieties to put the Chinese English data into a broader World Englishes perspective.

Fourthly, a linguistically motivated quality metric for evaluating the output of TTS systems was proposed (see chapter 5). The output of the TTS system we used previously had been evaluated either by visually comparing mel spectrograms or by mean opinion score evaluations. In line with the vowel space analysis methods used in the project, a novel metric based on the overlap of the complete vowel space and individual vowel areas was proposed.

Fifthly, the TTS system was put to the test in a pilot study of learning outcomes of participants speaking Czech and Italian English, with American English as a reference (see chapter 6). While the study did not show an improvement in learning outcomes, the non-native voices were consistently rated as warmer and more likable. A similar study with participants from China is ongoing. Even though the results do not support the initial hypothesis of the project, the findings that non-native synthetic voices are rated as warmer and more likable is a useful insight for all artificial intelligence and human-computer

interaction applications, depending on the desired application of the technology. As this field of inquiry is still in an early stage, since previous work has mostly focused on native varieties, more studies are necessary to eliminate possible idiosyncratic preferences and predispositions of the participants in this study[4].

In summary, the papers presented in this thesis advance the field of linguistics by (1) presenting a systematic review of features of Chinese English, (2) providing an in-depth, stratified description of the Chinese English vowel space of two previously researched, though less well stratified, and one previously unresearched province, and (3) showing Chinese English is indeed on the syllable-timed end of the speech rhythm spectrum. Furthermore, on the methodological side, the thesis advances the field of linguistics by (1) showing that vowel space analysis is a valid way to evaluate the quality of text-to-speech systems, (2) providing evidence that speech rhythm metrics seem to require a certain level of fluency to produce accurate results, and (3) that automated vowel formant analysis is a valid and reliable analysis method, again given a certain fluency of the speakers analyzed. Lastly, within the wider area of research on human-computer interaction in hybrid societies, the thesis shows that (1) it does not seem that non-native English varieties improve learning outcomes, even though more research to confirm these initial results is necessary, and (2) non-native synthetic voices are consistently rated as warmer and more likable than native voices.

7.3 Limitations

When drawing conclusions from research and when generalizing results, the limitations of the studies have to be kept in mind. The major limitations of the studies presented in this thesis fall into two categories: (1) data-related limitations; and (2) theoretical and methodological limitations.

The data collection process of the studies was severely impacted and delayed by the COVID-19 pandemic, as discussed in section 1.6.2. Due to the restrictions on international travel and on conducting in-person studies, the data collection took place later than initially planned and was also mostly conducted by colleagues at partner institutions. This led to some unanticipated effects such as instructions being interpreted in rather unexpected ways. Therefore, while trying to present stratified data in the sense of having equal numbers of participants in subgroups representing sociolinguistic variables, ie. equal numbers of female and male participants and equal numbers of participants from the Chinese provinces, the data is a convenience sample. Furthermore, since the data was collected at universities, it represents an educated variety. This well-known sampling bias is a common limitation of studies investigating non-native varieties, as proficiency in an L2 usually correlates with education. Conversely, this fact makes the data comparable, as this applies to many studies (see, for example, Ao and Low 2016; Deterding 2006b, 2016; Ho 2003; S. Li and Sewell 2012, all of which recruited participants at universities). On a related note, several recordings had to be excluded due to user errors while recording, leading to audio clipping or excessive microphone popping. These issues cannot be mitigated on the spot when data is not collected by the data analyst who knows what is important in the later stages of the project.

On the theoretical side, every study comes with an underlying set of assumptions that are important to consider when interpreting results. For example, when investigating L1 influence on L2 speech, as in chapter 3 and chapter 4, one of the underlying assumptions is that one's preferred Second Language

[4] A larger study is underway as part of the research project D03 (Tamboli, in prep.).

Acquisition theory allows L1 influence (cf. footnote 10 on page 12). Similarly, in the investigation of durational variation presented in chapter 4, the underlying assumption that measuring a surface phenomenon allows conclusions of underlying language features and cross-linguistic influence has to be accepted. Consequently, the results and conclusions presented in this thesis should be interpreted with the theoretical background and assumptions presented in chapter 1.

On the methodological side, isolating a variable and its effect in an experiment is challenging. In the case of the learning outcome study presented in chapter 6, measuring the effect of language cues on learning outcomes is difficult as there are numerous potential confounding variables. The study was designed to control for as many variables as possible by: (a) previous knowledge was controlled by selecting a topic that was on the one hand relevant to the participants' field of study but on the other hand previously unknown to them; (b) the language of the learning material that was adapted from a research article was assessed using several readability measures aimed at both native and non-native texts and was found to be appropriately challenging; (c) the material was kept relatively short to eliminate negative interference of attention span and memory capacity; (d) all synthetic voices had the same gender (female) to control socially conditioned bias; (e) the content was administered as audio only, without an accompanying transcript or other visual stimuli to avoid the visual channel as an influence on the learning outcomes; (f) the CPA was not given any appearance or personality, just the voice, to avoid evoking teacher stereotypes. As the study was conducted online, several aspects including the audio playback modality, whether speakers or headphones were used, and screen size (text flows differently on mobile devices and computers) could not be controlled.

In summary, the limitations imposed by external factors, such as a global pandemic, have been mitigated successfully. The methodological limitations result from the very nature of research in the humanities and have to be accounted for. In the end, a tradeoff between controlling confounding variables and the naturalness of an experiment has to be made. The compromise attempted here, especially in the work presented in chapter 6, was to keep the stimuli and the task as natural as possible while controlling as many variables as possible.

7.4 Future Research

Based on the general trajectory of the research and the limitations of the studies presented in this thesis, several promising directions for future research emerge. Future research that could be informed by the results of the papers presented here broadly falls into two categories: (1) linguistic research; and (2) research on credibility and learning.

On the linguistic side, the most important direction would be to include more data, both from the same provinces and from other parts of the country. It would be very interesting to see, how the North-South divide in Chinese varieties (cf. section 1.4) is mirrored in Chinese English. Furthermore, data from all parts of the country, possibly with more precise geolocation data could be used to create Chinese English language maps similar to maps in Siebenhaar and Hahn (2019). For this purpose, other data sources apart from recordings could be investigated, geotagged social media data might be a good source of data, even though availability and access would be a major challenge.

Including other sources of data, such as social media data, especially written data, would also allow quantitative research on features of Chinese English lexis and grammar. The features discussed in chapter 2, which mostly came from impressionistic studies, could serve as a starting point. Methodologically, automated analysis approaches to corpus data, including part-of-speech tagging could be used to analyze

large amounts of data, similar to the automated analysis of phonetic data presented in section 1.6.4. This approach could also be applied to social media data to include language from a less formal setting.

To investigate possible reasons for the poor reading performance of participants that was found in the data used in the studies presented in this thesis, a possible idea would be to investigate the acquisition of English grapheme-to-phoneme mapping by Chinese learners and to check if an effect of their L1 can be found. While studying L1 influence might be challenging due to the different writing system that Chinese languages use that has no direct correlation between spelling and pronunciation, the *pinyin* a transliteration system, which is used by most digital input methods on computers and smartphones, might be a good starting point, as it provides a direct grapheme-phoneme correlation.

On the credibility and learning side, using a validated scale to improve the comparability of results would be a logical next step. While such a scale is not yet available, it is currently being developed and validated as part of the work in the research project D03 (Tamboli, in prep.). Furthermore, the format of the stimuli presented in a survey could be varied to include different levels of formality, for example, the lecture style used in the study presented in chapter 6 is rather formal, so a podcast could be less formal while maintaining the audio-only mode. Additionally, a multimodal approach could be used, including slides to accompany the audio lecture, or using videos with the audio track replaced. While these approaches also introduce confounding variables, they could improve the naturalness of the situation.

Giving the CPA an appearance could also improve the naturalness of the situation and further intensify anthropomorphism processes and attribution of human characteristics to the agent. An interesting aspect that could be studied in such a scenario would be the mismatch of the agent's voice and appearance, for example, an agent with a caucasian appearance speaking with a Chinese English voice. This could potentially also add new insights into (language) stereotypes.

Lastly, on a broader scale, the text-to-speech system developed as part of the research presented here could be used to investigate the acceptance of non-native varieties in a wide range of scenarios, not limited to education.

Overall, this thesis presented new insights into regional variation in Chinese English vowel spaces and durational variation, as well as on durational variation of two underresearched varieties, Czech and Italian English. Furthermore, the study in chapter 5 has highlighted the fact that linguistic theories and methods can be applied successfully in computer science to create a linguistically motivated quality metric of monophthong naturalness. Finally, this thesis has provided evidence that considering non-native varieties in the design and implementation is a valid move to improve the favorable perception EDTs by their users.

8 Referencess

Abdulrahman, A., & Richards, D. (2022). Is Natural Necessary? Human Voice versus Synthetic Voice for Intelligent Virtual Agents. *Multimodal Technologies and Interaction*, *6*(7), 51. doi: 10.3390/mti6070051

Abelson, R. P. (1980). Searle's argument is just a set of chinese symbols. *Behavioral and Brain Sciences*, *3*, 424–425.

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Abu Guba, M. N., Mashaqba, B., Jarbou, S., & Al-Haj Eid, O. (2023). Production of vowel reduction by Jordanian–Arabic speakers of English: An acoustic study. *Poznan Studies in Contemporary Linguistics*, *59*(1), 1–25. doi: 10.1515/psicl-2022-2011

Adank, P., Smits, R., & van Hout, R. (2004). A Comparison of Vowel Normalization Procedures for Language Variation Research. *The Journal of the Acoustical Society of America*, *116*(5), 3099–3107. doi: 10.1121/1.1795335

Adank, P., van Heuven, V. J., & van Hout, R. (1999). Speaker Normalization Preserving Regional Accent Differences in Vowel Quality. In *Proceedings of the 14th international congress of phonetic sciences, san francisco* (pp. 1593–1596).

Ahn, J., & Moore, D. (2011). The relationship between students' accent perception and accented voice instructions and its effect on students' achievement in an interactive multimedia environment. *Journal of Educational Multimedia and Hypermedia*, *20*(4), 319–335.

Aiello, J. (2016). Accent, attitudes, and ownership of english: Perspectives of italian college-bound youth. In N. Tsantila, J. Mandalios, & I. Melpomeni (Eds.), *ELF: Pedagogical and interdisciplinary perspectives* (pp. 293–299). Deree-The American College of Greece.

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech* (pp. 103–113). London: Academic.

Albrecht, S. (2017). Corpus-Linguistic and Cognitive Approaches to Determiner Usage in Chinese Student Writing Testing the Fluctuation Hypothesis. In *Corpus Linguistics 2017 Conference.* University of Birmingham, 25-28 July 2017.

Albrecht, S. (2021). Current research on the linguistic features of Chinese English. *World Englishes*, weng.12572. doi: 10.1111/weng.12572

Albrecht, S. (2023a). *Analysis Scripts for "Durational Variation across inner/outer/expanding Circle Varieties of English".* Zenodo. doi: 10.5281/ZENODO.10041445

Albrecht, S. (2023b). *Analysis Scripts for "The Chinese English Vowel Space Across Three Provinces".* Zenodo. doi: 10.5281/ZENODO.8409882

Albrecht, S. (2023c). *Praat scripts for automated vowel formant measurements.* Zenodo. doi: 10.5281/ZENODO.8384068

Albrecht, S. (fc.). The Chinese English Vowel Space Across Three Provinces.

Albrecht, S., & Schmied, J. (2023). Comparing Vowel Spaces across Three Circles: A quantitative corpus-based Approach. In *IAWE25.*

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. In *2022 IEEE 9th International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)* (pp. 1–6). Chemnitz, Germany: IEEE. doi: 10.1109/CIVEMSA53371.2022.9853712

Algethami, G., & Hellmuth, S. (2023). Methods for investigation of L2 speech rhythm: Insights from the production of English speech rhythm by L2 Arabic learners. *Second Language Research*, 026765832311526. doi: 10.1177/02676583231152638

Alo, M., & Mesthrie, R. (2008). Nigerian English: Morphology and Syntax. In R. Mesthrie (Ed.), *Varieties of English. Africa, South and Southeast Asia* (Vol. 4, pp. 323–339). Berlin: Mouton de Gruyter Berlin.

Alvis, J. W. (2016). Indifference: Derrida Beyond Husserl, Intentionality, and Desire. In *Marion and Derrida on The Gift and Desire: Debating the Generosity of Things* (Vol. 85, pp. 125–153). Cham: Springer International Publishing. doi: 10.1007/978-3-319-27942-8_5

Andrist, S., Ziadee, M., Boukaram, H., Mutlu, B., & Sakr, M. (2015). Effects of culture on the credibility of robot speech: A comparison between english and arabic. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 157–164).

Antonelli, M. (2022). Consciousness and Intentionality in Franz Brentano. *Acta Analytica*, *37*(3), 301–322. doi: 10.1007/s12136-021-00480-2

Ao, R., & Low, E. L. (2012). Exploring pronunciation features of Yunnan English. *English Today*, *28*(3), 27–33.

Ao, R., & Low, E. L. (2016). A description of the Yunnan English accent. *World Englishes*, *35*(1), 18–41.

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics*, *24*(4), 621–635. doi: 10.1017/S0142716403000316

Arvaniti, A. (2009). Rhythm, Timing and the Timing of Rhythm. *Phonetica*, *66*(1-2), 46–63. doi: 10.1159/000208930

Arvaniti, A. (2012a). Rhythm classes and speech perception. In O. Niebuhr (Ed.), *Understanding prosody: The role of context, function and communication* (pp. 75–92). Berlin/Boston: De Gruyter.

Arvaniti, A. (2012b). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351–373. doi: 10.1016/j.wocn.2012.02.003

Arvaniti, A., & Rodriquez, T. (2013). The role of rhythm class, speaking rate, and F0 in language discrimination. *Laboratory Phonology*, *4*(1). doi: 10.1515/lp-2013-0002

Asimov, I. (1950). *I, robot*. New York: Doubleday.

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology*, *30*(1), 117–139. doi: 10.1016/j.cedpsych.2004.07.001

Baceviciute, S., Lucas, G., Terkildsen, T., & Makransky, G. (2022). Investigating the redundancy principle in immersive virtual reality environments: An eye-tracking and EEG study. *Journal of Computer Assisted Learning*, *38*(1), 120–136. doi: 10.1111/jcal.12595

Bailey, R. W., & Görlach, M. (Eds.). (1986). *English as a world language* (2. print ed.). Ann Arbor: Univ. of Michigan Press.

Bansal, S., & Aggarwal, C. (2022). *Textstat.* Computer software.

Barbosa, P. A. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication*, *49*(9), 725–742. doi: 10.1016/j.specom.2007.04.013

Barkana, B. D., & Patel, A. (2020). Analysis of vowel production in Mandarin/Hindi/American-accented English for accent recognition systems. *Applied Acoustics*, *162*, 107203. doi: 10.1016/j.apacoust.2019.107203

Bazzi, L., Brouwer, S., Planelles Almeida, M., & Foucart, A. (2022). Would you respect a norm if it sounds foreign? Foreign-accented speech affects decision-making processes. *PLOS ONE*, *17*(10), e0274727. doi: 10.1371/journal.pone.0274727

Beege, M., Schneider, S., Nebel, S., Mittangk, J., & Rey, G. D. (2017). Ageism – Age coherence within learning material fosters learning. *Computers in Human Behavior*, *75*, 510–519. doi: 10.1016/j.chb.2017.05.042

Behrman, A., Ferguson, S. H., Akhund, A., & Moeyaert, M. (2019). The Effect of Clear Speech on Temporal Metrics of Rhythm in Spanish-Accented Speakers of English. *Language and Speech*, *62*(1), 5–29. doi: 10.1177/0023830917737109

Bertinetto, P. M. (1981). *Strutture prosodiche dell'Italiano: Accento, quantità, sillaba, giuntura, fondamenti metrici* (Vol. 10). Presso l'Accademia della Crusca.

Bertinetto, P. M. (1989). Reflections on the dichotomy 'stress' vs. 'syllable-timing'. *Revue De Phonétique Appliquée*, *91*(93), 99–130.

Betts, R. F. (2006). *Decolonization* (2. ed., reprinted ed.). New York, NY: Routledge.

Bian, F. (2013). The Influence of Chinese Stress on English Pronunciation Teaching and Learning. *English Language Teaching*, *6*(11), 199–211.

Biber, D. (2003). Compressed noun-phrase structures in newspaper discourse: Thecompeting demands of popularization vs. economy. In J. Aitchison & D. Lewis (Eds.), *New Media Language* (pp. 169–181). London: Routledge.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Bione, T., & Cardoso, W. (2020). Synthetic voices in the foreign language context. *Language Learning & Technology*, *24*(1), 169–186. doi: 10.10125/44715

Block, N. (1980). What intuitions about homunculi don't show. *Behavioral and Brain Sciences*, *3*, 425–426.

Boduch-Grabka, K., & Lev-Ari, S. (2021). Exposing Individuals to Foreign Accent Increases their Trust in What Nonnative Speakers Say. *Cognitive Science*, *45*(11), e13064. doi: 10.1111/cogs.13064

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer*. Computer program.

Bolton, K. (2003). *Chinese Englishes: A Sociolinguistic History*. Cambridge University Press.

Bolton, K., & Graddol, D. (2012). English in China today. *English Today*, *28*(3).

Bolton, K., & Kachru, B. B. (2006). *World Englishes: Critical Concepts in Linguistics* (Vol. 2). Taylor & Francis.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* (4th ed. ed.). Berlin: Springer.

Brabcová, K., & Skarnitzl, R. (2018). Foreign or native-like? The attitudes of czech EFL learners towards accents of english and their use as pronunciation models. *Studie Z Aplikované Lingvistiky*, *1*, 38–50.

Braun, B., & Geiselmann, S. (2011). Italian in the no-man's land between stress-timing and syllable-timing?: Speakers are more stress-timed than listeners. In *Interspeech* (pp. 2697–2700).

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, *59*(1-2), 119–155. doi: 10.1016/S1071-5819(03)00018-1

Brentano, F. (1874). *Psychology from an empirical standpoint*. Abingdon, Oxon: Routledge.

Brentano, F. ([1874] 1973). *Psychologie vom empirischen Standpunkt. 1* (Unveränd. Nachdr. d. Ausg. von 1924 ed.; O. Kraus, Ed.). Hamburg: Meiner.

Brentano, F., & Hedwig, K. (1980). *Geschichte der mittelalterlichen Philosophie im christlichen Abendland* (No. Bd. 323). Hamburg: F. Meiner.

Bridgeman, B. (1980). Brains + programs = minds. *Behavioral and Brain Sciences*, *3*, 427–428.

Brinckmann, C. (2013). *Praat Scripting*. Oxford University Press. doi: 10.1093/oxfordhb/9780199571932.013.012

Brinckmann, C. (2017). Transcription bottleneck of speech corpus exploitation. In V. Lyding (Ed.), (pp. 165–179). Bozen: Europäische Akademie.

Brom, C., Hannemann, T., Stárková, T., Bromová, E., & Děchtěrenko, F. (2017). The role of cultural background in the personalization principle: Five experiments with Czech learners. *Computers & Education*, *112*, 37–68. doi: 10.1016/j.compedu.2017.01.001

Brown, G. (2013). *Listening to spoken english (2nd)*. Routledge.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Buck, G. (2007). *Assessing Listening* (1st ed.). Cambridge University Press. doi: 10.1017/CBO9780511732959

Butler, S. (1997). Corpus of English in Southeast Asia: Implications for a Regional Dictionary. *English is an Asian language: The Philippine context*, 103–124.

Cannon, G. (1988). Chinese Borrowings in English. *American Speech*, *63*(1), 3–33.

Capek, J. (1925). *Lelio A Pro Delfina*. Prague: Aventinum.

Capek, K. (1920). *R.U.R. (Rossum's Universal Robots)*. Prague: Aventinum.

Cargile, A. C., & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language & Communication*, *17*(3), 195–217. doi: https://doi.org/10.1016/S0271-5309(97)00016-5

Caston, V. (1998). Aristotle and the Problem of Intentionality. *Philosophy and Phenomenological Research*, *58*(2), 249. doi: 10.2307/2653509

Caston, V. (in press). The stoics on mental representation. *The Oxford Handbook in Hellenistic Philosophy, Oxford*.

Chambers, J. K. (2009). *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Oxford: Wiley-Blackwell.

Chambers, J. K., & Trudgill, P. (1998). *Dialectology* (2nd ed.). Cambridge University Press. doi: 10.1017/CBO9780511805103

Chan, H. L., & Fuchs, R. (fc.). Revisiting the Vowels of Hong Kong English - The post-handover Generation. In *ICPhS*.

Chan, K. Y., & Hall, M. D. (2019). The importance of vowel formant frequencies and proximity in vowel space to the perception of foreign accent. *Journal of Phonetics*, *77*, 100919. doi: 10.1016/j.wocn.2019.100919

Chang, J. (1987). Chinese Speakers. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (1st ed., pp. 224–237). Cambridge: Cambridge University Press.

Chang, J. (2001). Chinese Speakers. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (2nd ed., pp. 310–324). Cambridge: Cambridge University Press.

Chao, Y. R. (1968). *A Grammar of Spoken Chinese.* (2nd ed.). Berkeley: University of California Press.

Chao, Y. R. (1980). *Language and symbolic systems* (Repr ed.). Cambridge: UP.

Chen, G. T. (1972). *A comparative study of pitch range of native speakers of Midwestern English and Mandarin Chinese: An acoustic study*. University of Wisconsin.

Chen, H. C., & Wang, M. J. (2011). An acoustic analysis of Chinese and English vowels. *CALR Linguistic Journal*(1), 1–19.

Chen, P. (2003). Indefinite determiner introducing definite referent: A special use of 'yi 'one'+ classifier'in Chinese. *Lingua*, *113*(12), 1169–1184.

Chen, P. (2004). Identifiability and definiteness in Chinese. *Linguistics*, *42*(6), 1129–1184.

Chen, Y. (2008). The acoustic realization of vowels of Shanghai Chinese. *Journal of Phonetics*, *36*(4), 629–648. doi: 10.1016/j.wocn.2008.03.001

Chen, Y., & Gussenhoven, C. (2015). Shanghai Chinese. *Journal of the International Phonetic Association*, *45*(3), 321–337. doi: 10.1017/S0025100315000043

Chen, Y., & Robb, M. (2000). Acoustic features of vowel production in Mandarin speakers of English. In *Sixth international conference on spoken language processing.*

Chen, Y., Robb, M., Gilbert, H., & Lerman, J. (2001). Vowel production by Mandarin speakers of English. *Clinical Linguistics & Phonetics*, *15*(6), 427–440.

Cheng, A., Faytak, M., & Cychosz, M. (2016). Language, race, and vowel space: Contemporary Californian English. In *Proceedings of the forty-second annual meeting of the berkeley linguistics society* (pp. 63–78).

Cheng, C.-C. (1992). Chinese Varieties of English. In B. B. Kachru (Ed.), *The Other Tongue: English across Cultures* (2nd ed., pp. 162–177). Urbana, IL: University of Illinois Press.

Cheng, S. W. (2005). *An exploratory cross-sectional study of interlanguage pragmatic development of expressions of gratitude by Chinese learners of English* (Unpublished doctoral dissertation). University of Iowa.

Cheshire, J. (2004). Sex and Gender in Variationist Research. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (1st ed., pp. 423–443). Wiley. doi: 10.1002/9780470756591.ch17

Cheshire, J. (2005). Age- and Generation-Specific Use of Language. In U. Ammon, N. Dittmar, K. J. Mattheier, & P. Trudgill (Eds.), *Sociolinguistics* (pp. 1552–1563). Walter de Gruyter. doi: 10.1515/9783110171488.2.8.1552

Cheung, W. H. (2009). Span of High Tones in Hong Kong English. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 35, pp. 72–82).

Chevelu, J., Lolive, D., Maguer, S. L., & Guennec, D. (2015). How to compare TTS systems: A new subjective evaluation methodology focused on differences. In *Sixteenth annual conference of the international speech communication association.*

Chisholm, R. M. (1990). Brentano and Marty on Content: A Synthesis Suggested by Brentano. In K. Mulligan (Ed.), *Mind, Meaning and Metaphysics* (pp. 1–9). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-009-0505-4_1

Cho, E., Sundar, S. S., Abdullah, S., & Motalebi, N. (2020). Will Deleting History Make Alexa More Trustworthy?: Effects of Privacy and Content Customization on User Experience of Smart

Speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Honolulu HI USA: ACM. doi: 10.1145/3313831.3376551

Choe, W. K. (2019). The realization of English rhythm by Busan Korean speakers*. *Phonetics and Speech Sciences*, *11*(4), 81–87. doi: 10.13064/KSSS.2019.11.4.081

Chomsky, N. (1957). *Syntactic Structures*. De Gruyter. doi: 10.1515/9783112316009

Chomsky, N. (1965). *Aspects of the theory of syntax* (50th ed.). The MIT Press.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Christensen, M., Welch, A., & Barr, J. (2017). Husserlian Descriptive Phenomenology: A review of intentionality, reduction and the natural attitude. *Journal of Nursing Education and Practice*, *7*(8), 113. doi: 10.5430/jnep.v7n8p113

Chu, M., & Peng, H. (2006). *Objective measure for estimating mean opinion score of synthesized speech.* Google Patents. (US Patent 7,024,362)

Chuang, F.-Y. (2005). Article misuse: A neglected problem in Chinese EAP student writing. In *Proceedings of the 2nd international online conference on second and foreign language teaching and research* (pp. 25–33).

Churchland, P., & Churchland, P. S. (1990). Could a machine think? *Scientific American*, *262*(1), 32–39.

Clopper, C. G. (2009). Computational Methods for Normalizing Acoustic Vowel Data for Talker Differences: Computational Methods for Normalizing Acoustic Vowel Data. *Language and Linguistics Compass*, *3*(6), 1430–1442. doi: 10.1111/j.1749-818X.2009.00165.x

Cole, D. (1991). Artificial intelligence and personal identity. *Synthese. An International Journal for Epistemology, Methodology and Philosophy of Science*, *88*, 399–417.

Copeland, B. J. (1993). *Artificial intelligence: A philosophical introduction*. Blackwell.

Cornford, F. M. (2013). *Plato's Theory of Knowledge* (??th ed.). Routledge. doi: 10.4324/9781315822884

Craig, E. (1996). *Routledge encyclopedia of philosophy: Genealogy to iqbal*. New York: Routledge.

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, *114*, 193–205. doi: 10.1016/j.compedu.2017.07.003

Craig, S. D., & Schroeder, N. L. (2019). Text-to-Speech Software and Learning: Investigating the Relevancy of the Voice Effect. *Journal of Educational Computing Research*, *57*(6), 1534–1548. doi: 10.1177/0735633118802877

Crane, T. (1998). Intentionality as the Mark of the Mental. *Royal Institute of Philosophy Supplement*, *43*, 229–251. doi: 10.1017/S1358246100004380

Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., & Floccia, C. (2012). Linguistic Processing of Accented Speech Across the Lifespan. *Frontiers in Psychology*, *3*. doi: 10.3389/fpsyg.2012.00479

Crystal, D. (2003). *English as a Global Language* (2nd ed.). Cambridge University Press. doi: 10.1017/CBO9780511486999

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Syllabic stress. *The Journal of the Acoustical Society of America*, *83*(4), 1574–1585. doi: 10.1121/1.395912

Cuerie, H. C. (1952). A projection of socio-linguistics: The relationship of speech to social status. *The Southern Speech Journal*, *18*(1), 28–37. doi: 10.1080/10417945209371247

Cui, R., & van Heuven, V. J. (2011). Mutual intelligibility of English vowels by Chinese dialect speakers.

In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th international congress of phonetic sciences* (pp. 544–547).
Cummins, F. (2002). Speech rhythm and rhythmic taxonomy. In *Proceedings of speech prosody 2002* (pp. 121–126). Aix-en-Provence.
Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*(2), 145–171. doi: 10.1006/jpho.1998.0070
Cutler, A. (2012). *Native Listening: Language Experience and the Recognition of Spoken Words*. The MIT Press. doi: 10.7551/mitpress/9012.001.0001
Cutler, A., & Van Donselaar, W. (2001). Voornaam is not (really) a Homophone: Lexical Prosody and Lexical Access in Dutch. *Language and Speech*, *44*(2), 171–195. doi: 10.1177/00238309010440020301
Dahlbäck, N., Swamy, S., Nass, C., Arvidsson, F., & Skågeby, J. (2001). Spoken interaction with computers in a native or non-native language - same or different? In *Proceedings of INTERACT.*
Dahlbäck, N., Wang, Q., Nass, C., & Alwin, J. (2007). Similarity is more important than expertise: Accent effects in speech interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1553–1556).
Dai, L., Jung, M. M., Postma, M., & Louwerse, M. M. (2022). A systematic review of pedagogical agent research: Similarities, differences and unexplored aspects. *Computers & Education*, *190*, 104607. doi: 10.1016/j.compedu.2022.104607
Dankovičová, J. (1997). The domain of articulation rate variation in Czech. *Journal of Phonetics*, *25*(3), 287–312. doi: 10.1006/jpho.1997.0045
Dankovičová, J., & Dellwo, V. (2007). Czech speech rhythm and the rhythm class hypothesis. doi: 10.5167/UZH-111796
Danto, A. C. (1980). The use and mention of terms and the simulation of linguistic understanding. *Behavioral and Brain Sciences*, *3*, 428.
Darcy, I., & Thomas, T. (2019). When blue is a disyllabic word: Perceptual epenthesis in the mental lexicon of second language learners. *Bilingualism*, *22*(5), 1141–1159.
Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62. doi: 10.1016/S0095-4470(19)30776-4
Davis, R. O., Vincent, J., & Park, T. (2019). Reconsidering the Voice Principle with Non-native Language Speakers. *Computers & Education*, *140*, 103605. doi: 10.1016/j.compedu.2019.103605
de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. doi: 10.3758/BRM.41.2.385
Défossez, A., Synnaeve, G., & Adi, Y. (2022). *Denoiser.* Computer software.
Delibegović Džanić, N., & Berberović, S. (2021). Lemons and Watermelons: Visual Advertising and Conceptual Blending. In L. D'Angelo, A. Mauranen, & S. Maci (Eds.), *Metadiscourse in Digital Communication* (pp. 115–132). Cham: Springer International Publishing. doi: 10.1007/978-3-030-85814-8_6
Dellwo, V. (2006). Rhythm and Speech Rate: A Variation Coefficient for deltaC. doi: 10.5167/UZH-111789
Dellwo, V. (2010). *Influences of speech rate on the acoustic correlates of speech rhythm : An experimental phonetic study based on acoustic and perceptual evidence* (Unpublished doctoral dissertation). Rheinische Friedrich-Wilhelms-Universität Bonn.

Dellwo, V., & Wagner, P. (2003). Relations between language rhythm and speech rate. doi: 10.5167/UZH-111779

Demirsahin, I., Kjartansson, O., Gutkin, A., & Rivera, C. (2020). Open-Source Multi-Speaker Corpora of the English Accents in the British Isles. In *Proceedings of the 12th language resources and evaluation conference (LREC)* (pp. 6532–6541). Marseille, France: European Language Resources Association (ELRA).

Dennett, D. C. (1980). The milk of human intentionality. *Behavioral and Brain Sciences*, *3*, 428–430.

Dennett, D. C. (2010). *Content and Consciousness* (1st ed.). Routledge. doi: 10.4324/9780203092958

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. New York: W.W. Norton and Co.

Descartes, R. (1641). *Meditationes de prima philosophia*.

Deterding, D. (2006a). 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.

Deterding, D. (2006b). The Pronunciation of English by Speakers from China. *English World-Wide*, *27* (2), 175–198.

Deterding, D. (2016). The Pronunciation of English by Students in Guangxi, South China. *rEFLections*, *22*, 1–20.

Deterding, D. (2017). The Pronunciation of English in Guangxi: Which Features Cause Misunderstandings? In Z. Xu, D. He, & D. Deterding (Eds.), *Researching Chinese English: The State of the Art* (pp. 17–31). Springer.

Deterding, D., Wong, J., & Kirkpatrick, A. (2008). The Pronunciation of Hong Kong English. *English World-Wide*, *29*(2), 148–175.

Ding, H., Hoffmann, R., & Hirst, D. (2016). Prosodic Transfer: A Comparison Study of F0 Patterns in L2 English by Chinese Speakers. In *Speech Prosody* (Vol. 2016, pp. 756–760).

Ding, H., Lin, B., Wang, L., Wang, H., & Fang, R. (2020). A Comparison of English Rhythm Produced by Native American Speakers and Mandarin Esl Primary School Learners. In *INTERSPEECH* (pp. 4481–4485).

Ding, H., & Xu, X. (2016). L2 english rhythm in read speech by chinese students. In *INTERSPEECH* (pp. 2696–2700).

Di Paolo, M., Yaeger-Dror, M., & Wassink, A. B. (2011). Analyzing vowels. In M. Di Paolo & M. Yaeger-Dror (Eds.), *Sociophonetics: A student's guide* (pp. 87–106). New York: Routledge.

Disner, S. F. (1980). Evaluation of vowel normalization procedures. *The Journal of the Acoustical Society of America*, *67* (1), 253–261. doi: 10.1121/1.383734

Do, T. D., Akter, M., Choudhary, Z., Azevedo, R., & McMahan, R. P. (2022). The Effects of an Embodied Pedagogical Agent's Synthetic Speech Accent on Learning Outcomes. In *International Conference on Multimodal Interaction* (pp. 198–206). Bengaluru India: ACM. doi: 10.1145/3536221.3556587

Dolson, M. (1994). The Pitch of Speech as a Function of Linguistic Community. *Music Perception*, *11*(3), 321–331.

Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing*. London: Lawrence Erlbaum.

Drager, K. (2010). Sociophonetic Variation in Speech Perception. *Language and Linguistics Compass*, *4*(7), 473–480. doi: 10.1111/j.1749-818X.2010.00210.x

Dreyfus, H. L. (1972). *What computers can't do: The limits of artificial intelligence*. New York: MIT Press.

Dreyfus, H. L., & Dreyfus, S. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. Oxford, U.K.: Blackwell.
Du, R., & Jiang, Y. (2001). Jin ershi nian "Zhongguo yingyu" yanjiu shuping ["China English" in the Past 20 Years]. *Waiyu Jiaoxue Yu Yanjiu (Foreign Language Teaching and Researching)*, *33*(1), 37–41.
Du, Z. (1998). Zhongguo yingyu wenti ji qita [Issues of Chinese English and others]. *Waiyu jiaoxue [Foreign Langauge Education]*, *19*(3), 6–14.
Duanmu, S. (2007). *The Phonology of Standard Chinese*. Oxford: Oxford University Press.
Duběda, T. (2004). K izosylabičnosti a izochronnosti v Češtině. In *Proceedings of the conference of the Czech section of ISPhS* (pp. 19–28).
Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3-4), 177–190. doi: 10.1016/S0921-8890(02)00374-3
Du Toit, J., & Verhoef, A. H. (2018). Embodied digital technology and transformation in higher education. *Transformation in Higher Education*, *3*. doi: 10.4102/the.v3i0.52
Eady, S. J. (1982). Differences in the F0 Patterns of Speech: Tone Language Versus Stress Language. *Language and speech*, *25*(1), 29–42.
Eaves, M. (2011). English, Chinglish or China English?: Analysing Chinglish, Chinese English and China English. *English Today*, *27*(4), 64–70.
Eberhard, D. M., Simons, G. F., Fennig, C. D., & Summer Institute of Linguistics (Eds.). (2022). *Ethnologue: Languages of the World* (Twenty-sixth edition ed.). Dallas, Texas: SIL International.
Eccles, J. C. (1980). A dualist-interactionist perspective. *Behavioral and Brain Sciences*, *3*, 430–431.
Eckert, P. (1988). Adolescent social structure and the spread of linguistic change. *Language in Society*, *17*(2), 183–207. doi: 10.1017/S0047404500012756
Eckert, P. (2000). *Language Variation as Social Practice*. Oxford and Malden: Blackwell.
Eckert, P., & McConnell-Ginet, S. (1999). New generalizations and explanations in language and gender research. *Language in Society*, *28*(02). doi: 10.1017/S0047404599002031
Edwards, A., Edwards, C., Spence, P. R., Harris, C., & Gambino, A. (2016). Robots in the classroom: Differences in students' perceptions of credibility and learning between "teacher as robot" and "robot as teacher". *Computers in Human Behavior*, *65*, 627–634. doi: 10.1016/j.chb.2016.06.005
Ehret, J., Bönsch, A., Aspöck, L., Röhr, C. T., Baumann, S., Grice, M., ... Kuhlen, T. W. (2021). Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech? *ACM Transactions on Applied Perception*, *18*(4), 1–15. doi: 10.1145/3486580
Eisenstein, M., & Bodman, J. W. (1986). 'I very appreciate': Expressions of gratitude by native and non-native speakers of American English. *Applied linguistics*, *7*(2), 167–185.
Engwall, O., Lopes, J., & Åhlund, A. (2021). Robot Interaction Styles for Conversation Practice in Second Language Learning. *International Journal of Social Robotics*, *13*(2), 251–276. doi: 10.1007/s12369-020-00635-y
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. doi: 10.1037/0033-295X.114.4.864
Epstein, S. D., Flynn, S., & Martohardjono, G. (1996). Second language acquisition: Theoretical and experimental issues in contemporary research. *Behavioral and Brain Sciences*, *19*(4), 677–714.
Epstein, S. D., Flynn, S., & Martohardjono, G. (1998). The strong continuity hypothesis: Some evidence concerning functional categories in adult L2 acquisition. *The Generative Study of*

Second Language Acquisition.
Erickson, T. (1997). *Designing agents as if people mattered.* http://www.pliant.org/personal/Tom_Erickson/agents.html.
Eriksson, A. (1991). *Aspects of swedish speech rhythm* (Unpublished doctoral dissertation). Göteborg.
Evans, J. R., & Michael, S. W. (2014). Detecting Deception in Non-Native English Speakers: Detecting deception in non-native speakers. *Applied Cognitive Psychology*, *28*(2), 226–237. doi: 10.1002/acp.2990
Fabricius, A. H., Watt, D., & Johnson, D. E. (2009). A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change*, *21*(3), 413–435. doi: 10.1017/S0954394509990160
Feng, X. (2013). Zhongguo Yingyu'cunzai de helixing yu yifa jianlun [The rationality of the existence of Chinese English and its translating methods]. *Qilu shifan xueyuan xuebao [Journal of Qilu Normal University]*, *28*(5), 125–129.
Ferragne, E., & Pellegrino, F. (2004). A comparative account of the suprasegmental and rhythmic features of British English dialects. In *MIDL 2004.* Paris, France.
Finkel, M., & Krämer, N. C. (2022). Humanoid Robots – Artificial. Human-like. Credible? Empirical Comparisons of Source Credibility Attributions Between Humans, Humanoid Robots, and Non-human-like Devices. *International Journal of Social Robotics*, *14*(6), 1397–1411. doi: 10.1007/s12369-022-00879-w
Finn, A. N., Schrodt, P., Witt, P. L., Elledge, N., Jernberg, K. A., & Larson, L. M. (2009a). A meta-analytical review of teacher credibility and its associations with teacher behaviors and student outcomes. *Communication education*, *58*(4), 516–537.
Finn, A. N., Schrodt, P., Witt, P. L., Elledge, N., Jernberg, K. A., & Larson, L. M. (2009b). A Meta-Analytical Review of Teacher Credibility and its Associations with Teacher Behaviors and Student Outcomes. *Communication Education*, *58*(4), 516–537. doi: 10.1080/03634520903131154
Fischer, J. L. (1958). Social Influences on the Choice of a Linguistic Variant. *Word*, *14*(1), 47–56. doi: 10.1080/00437956.1958.11659655
Fletcher, J. (2010). The Prosody of Speech: Timing and Rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (1st ed., pp. 521–602). Wiley. doi: 10.1002/9781444317251.ch15
Flynn, N., & Foulkes, P. (2011). Comparing vowel formant normalization methods..
Fodor, J. A. (1980). Searle on what only brains can do. *Behavioral and Brain Sciences*, *3*, 431–432.
Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99* (pp. 80–87). Pittsburgh, Pennsylvania, United States: ACM Press. doi: 10.1145/302979.303001
Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General*, *112*(3), 386–412. doi: 10.1037/0096-3445.112.3.386
Fox, R. A., & Jacewicz, E. (2017). Reconceptualizing the vowel space in analyzing regional dialect variation and sound change in American English. *The Journal of the Acoustical Society of America*, *142*(1), 444–459. doi: 10.1121/1.4991021
Fuchs, R. (2016). Speech Rhythm in Varieties of English. In *Speech Rhythm in Varieties of English* (pp. 87–102). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-662-47818-9_4
Fuchs, R. (2023a). 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.* Amsterdam: Benjamins.

Fuchs, R. (2023b). A synthesis of research on speech rhythm in native, learner and second language varieties of english – introduction to the volume. In R. Fuchs (Ed.), *Speech rhythm in learner and second language varieties of english.* Singapore: Springer.

Gahl, S., & Baayen, R. H. (2019). Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood. *Journal of Phonetics*, *74*, 42–54. doi: 10.1016/j.wocn.2019.02.001

Ge, C. (1980). Mantan you han yi ying wenti (Random thoughts on some problems in Chinese-English translation). *Fanyi Tongxun (Chinese Translator's Journal)*, *2*, 1–8.

Ge, C. (1983). *Translation Theory and Translation Skills*. Beijing: China Translation Publishing Company.

Gera, D. L. (2003). *Ancient Greek ideas on speech, language, and civilization*. Oxford: Oxford University Press.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*(1), 78–80. doi: 10.1109/TAU.1968.1161953

Ghitza, O. (2011). Linking Speech Perception and Neurophysiology: Speech Decoding Guided by Cascaded Oscillators Locked to the Input Rhythm. *Frontiers in Psychology*, *2*. doi: 10.3389/fpsyg.2011.00130

Gill, M. M. (1994). Accent and stereotypes: Their effect on perceptions of teachers and lecture comprehension. *Journal of Applied Communication Research*, *22*(4), 348–361. doi: 10.1080/00909889409365409

Görlach, M. (1990). *Studies in the History of the English Language*. Heidelberg: Carl Winter.

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193.

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. doi: 10.1515/9783110197105.2.515

Gu, G., & Xiang, M. (1997). Wei zhongguo yingyu yibian [An argument for China English]. *Waiyu yu Waiyu Jiaoxue [Foreign languages and foreign languages teaching]*, *5*, 11–13.

Gu, Y. (1990). Politeness Phenomena in Modern Chinese. *Journal of Pragmatics*, *14*(2), 237–257.

Gupta, A. F. (1997). Colonisation, migration, and functions of English. In E. W. Schneider (Ed.), *Varieties of English Around the World* (Vol. G18, p. 47). Amsterdam: John Benjamins Publishing Company. doi: 10.1075/veaw.g18.08gup

Gussenhoven, C. (2012). Tone and intonation in Cantonese English. In *Tonal Aspects of Languages-Third International Symposium.*

Gut, U. (2005). Nigerian English prosody. *English World-Wide*, *26*(2), 153–177. doi: 10.1075/eww.26.2.03gut

Gut, U. (2008). Nigerian English: Phonology. In R. Mesthrie (Ed.), *Varieties of English. Africa, South and Southeast Asia* (Vol. 4, pp. 35–54). Berlin: Mouton de Gruyter Berlin.

Habermas, J. (1981). *Theorie des kommunikativen Handelns*. Frankfurt a.M.: suhrkamp.

Haddican, B., Foulkes, P., Hughes, V., & Richards, H. (2013). Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change*, *25*(3), 371–403. doi: 10.1017/S0954394513000197

Halák, J. (2023). Embodied higher cognition: Insights from Merleau-Ponty's interpretation of motor intentionality. *Phenomenology and the Cognitive Sciences*, *22*(2), 369–397. doi: 10.1007/s11097-021-09769-4

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

Halliday, M. A. K., McIntosh, A., & Strevens, P. (1964). *The Linguistic Sciences and Language Teaching*. London: Longmans.

Hampel, F. R. (1974). 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

Hanzlíková, D., & Skarnitzl, R. (2017). Credibility of native and non-native speakers of English revisited: Do non-native listeners feel the same? *Research in Language*, *15*(3), 285–298. doi: 10.1515/rela-2017-0016

Haspelmath, M. (2012). How to compare major word-classes across the world's languages. *Theories of Everything: In Honor of Edward Keenan*, *17*, 109–130.

Haugeland, J. (1980). Programs, causal powers, and intentionality. *Behavioral and Brain Sciences*, *3*, 432–433.

Hawkins, R., & Chan, C. Y.-h. (1997). The partial availability of Universal Grammar in second language acquisition: The 'failed functional features hypothesis'. *Second language research*, *13*(3), 187–226.

Hazen, K. (2010). Labov: Language Variation and Change. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics.* SAGE.

He, D., & Li, D. C. (2009). Language attitudes and linguistic features in the 'China English' debate. *World Englishes*, *28*(1), 70–89.

Heidegger, M. ([1927] 2006). *Sein und Zeit* (19. Aufl., unveränd. Nachdr. d. 15. Aufl ed.). Tübingen: Niemeyer.

Hilbert, M. (2020). Digital technology and social change: The digital transformation of society from a historical perspective. *Dialogues in Clinical Neuroscience*, *22*(2), 189–194. doi: 10.31887/DCNS.2020.22.2/mhilbert

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. doi: 10.1121/1.411872

Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. In D. Sankoff (Ed.), *Linguistic variation: Models and methods* (pp. 161–171). New York: Academic Press.

Ho, L. (2003). Pronunciation Problems of PRC Students. In G. L. Lee, L. Ho, J. L. Meyer, & C. Varaprasad (Eds.), *Teaching English to students from China* (pp. 138–157). Singapore: Singapore University Press Singapore.

Hockett, C. F. (1958). *A Course in Modern Linguistics*. New York: Macmillan Press.

Hockstein, N. G., Gourin, C. G., Faust, R. A., & Terris, D. J. (2007). A history of robots: From science fiction to surgical robotics. *Journal of Robotic Surgery*, *1*(2), 113–118. doi: 10.1007/s11701-007-0021-2

Hoffman, M. F., & Walker, J. A. (2010). Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change*, *22*(1), 37–67. doi: 10.1017/S0954394509990238

Hoijer, H. (1971). The sapir-whorf hypothesis. In H. Hoijer (Ed.), *Language in culture* (7th Impression

ed., pp. 92–105).
Holliday, N. (2023). Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication*, *8*, 1116955. doi: 10.3389/fcomm.2023.1116955
Holm, J. (2000). *An Introduction to Pidgins and Creoles* (1st ed.). Cambridge University Press. doi: 10.1017/CBO9781139164153
Honderich, T. (1995). *The oxford companion to philosophy* (T. Honderich, Ed.). New York: Oxford University Press.
Horvath, B. M. (1985). *Variation in Australian English: The Sociolects of Sydney* (No. 45). Cambridge: Cambridge University Press.
Hu, X. (2011). Zhongguo Yingyu bianti zhi jiaoji celue gongneng yanjiu [A study on the function of communicative strategy of the Chinese varieties of English]. *Liaoning shifan daxue xuebao [Journal of Liaoning Normal University]*, *34*(3), 103–106.
Huang, C.-T. J. (1987). Existential sentences in Chinese and (in) definiteness. *The representation of (in) definiteness*, *226*, 253.
Huang, J. (1988). Yingdang kending 'xiyi hanhua'xianxiang de jijimian (The positive role of Sinicism in the Englishtranslated version). *Zhongguo fanyi [Chinese Translators Journal]*, *1*, 39–47.
Hundt, M., & Mair, C. (1999). " Agile" and" Uptight" Genres: The Corpus-Based Approach to Language Change in Progress. *International Journal of Corpus Linguistics*, *4*(2), 221–242.
Hung, T. T. (2005). Phonological features of 'New englishes'. In *1st International Conference on the Linguistics of Contemporary English.* Edinburgh.
Hymes, D. (1968). The Ethnography of Speaking. In J. A. Fishman (Ed.), *Readings in the Sociology of Language* (pp. 99–138). De Gruyter. doi: 10.1515/9783110805376.99
Hymes, D. (Ed.). (1974). *Foundations in Sociolinguistics* (??th ed.). Philadelphia: University of Pennsylvania Press. doi: 10.4324/9781315888835
Ijalba, E., & Obler, L. K. (2015). First language grapheme-phoneme transparency effects in adult second-language learning.
Ionin, T. (2003). *Article Semantics in Second Language Acquisition* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
Ionin, T., Ko, H., & Wexler, K. (2004). Article Semantics in L2 Acquisition: The Role of Specificity. *Language Acquisition*, *12*(1), 3–69.
Isogai, M., Mizuno, H., & Mano, K. (2005). Recording Script Design for Corpus-Based TTS System Based on Coverage of Various Phonetic Elements. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* (Vol. 1, pp. 301–304). Philadelphia, Pennsylvania, USA: IEEE. doi: 10.1109/ICASSP.2005.1415110
Ito, K., & Johnson, L. (2017). *The LJ Speech Dataset.* `https://keithito.com/LJ-Speech-Dataset`.
Jaffee, D. (1997). Asynchronous Learning: Technology and Pedagogical Strategy in a Distance Learning Course. *Teaching Sociology*, *25*(4), 262. doi: 10.2307/1319295
Jenkins, J. (2012). *The Phonology of English as an International Language: New Models, New Norms, New Goals* (Nachdr. ed.). Oxford: Oxford University Press.
Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*. doi: 10.1007/s10639-023-11834-1
Jiang, W. (2017). A Study on Modified-Modifying Sequence in the Compositions by Chinese Advanced Users of English. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching Chinese English: The*

State of the Art (pp. 93–107). Springer.
Jiang, Y. (2002). China English: Issues, Studies and Features. *Asian Englishes*, *5*(2), 4–23.
Jimenez-Molina, A., Retamal, C., & Lira, H. (2018). Using Psychophysiological Sensors to Assess Mental Workload During Web Browsing. *Sensors*, *18*(2), 458. doi: 10.3390/s18020458
Joseph, J. E. (2004). *Language and Identity: National, Ethnic, Religious*. Basingstoke: Palgrave Macmillan.
Kachru, B. B. (1965). The Indianness in Indian English. *Word*, *21*(3), 391–410. doi: 10.1080/00437956.1965.11435436
Kachru, B. B. (1984). World Englishes and the teaching of English to non-native speakers: Contexts, attitudes, and concerns. *TESOL Newsletter*, *18*(5), 25–26.
Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: English language in the outer circle. In R. Quirk Randolph & H. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–36). Cambridge: Cambridge University Press.
Kachru, B. B. (1992a). Models for non-native Englishes. In B. B. Kachru (Ed.), *The Other Tongue: English across Cultures* (2nd ed., pp. 48–74). Urbana: University of Illinois Press.
Kachru, B. B. (1992b). World Englishes: Approaches, issues and resources. *Language Teaching*, *25*(1), 1–14. doi: 10.1017/S0261444800006583
Kachru, B. B. (1996). World Englishes: Agony and Ecstasy. *Journal of Aesthetic Education*, *30*(2), 135–155.
Kachru, B. B. (1998). English as an Asian language. *Links & letters*, 89–108.
Kant, I. (1781). *Critik der reinen vernunft*. Riga.
Karakaş, A. (2017). English voices in 'Text-to-speech tools': Representation of English users and their varieties from a World Englishes perspective. *Advances in Language and Literary Studies*, *8*(5), 108. doi: 10.7575/aiac.alls.v.8n.5p.108
Kartal, G. (2010). Does language matter in multimedia learning? Personalization principle revisited. *Journal of Educational Psychology*, *102*(3), 615–624. doi: 10.1037/a0019345
Kassambara, A. (2023a). *Ggpubr: 'ggplot2' based publication ready plots* [Manual].
Kassambara, A. (2023b). *Rstatix: Pipe-friendly framework for basic statistical tests* [Manual].
Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050–1060.
Kelly, M. R. (2016). The Truly Transcendental: Merleau-Ponty, un Écart, "The Acceptance of the Truth of the Transcendental Analysis". In *Phenomenology and the Problem of Time* (pp. 139–174). London: Palgrave Macmillan UK. doi: 10.1057/978-1-137-31447-5_5
Kinoshita, N., & Sheppard, C. (2011). Validating acoustic measures of speech rhythm for second language acquisition. In *ICPhS* (Vol. 17, pp. 1086–1089).
Kirkpatrick, A. (1993). Information sequencing in Modern Standard Chinese in a genre of extended spoken discourse. *Text - Interdisciplinary Journal for the Study of Discourse*, *13*(3). doi: 10.1515/text.1.1993.13.3.423
Kirkpatrick, A. (1996). 'Topic — Comment or Modifier — Modified? Information Structure in Modern Standard Chinese. *Studies in Language*, *20*(1), 93–113. doi: 10.1075/sl.20.1.05kir
Kirkpatrick, A. (2007). *World Englishes Hardback with Audio Cd: Implications for International Communication and English Language Teaching*. Cambridge University Press.
Kirkpatrick, A., & Xu, Z. (2002). Chinese pragmatic norms and 'China English'. *World Englishes*, *21*(2), 269–279.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221. doi: 10.1121/1.380986

Klautau, A. (2001). *ARPABET and the TIMIT alphabet.* http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf.

Kleinschmidt, D. F., Weatherholtz, K., & Florian Jaeger, T. (2018). Sociolinguistic Perception as Inference Under Uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834. doi: 10.1111/tops.12331

Knowlton, E. C. (1970). Chinese, Japanese, and Korean Loanwords in Webster's Third. *American Speech*, *45*(1/2), 8–29.

Kohring, M. (2001). Vertrauen in Medien - Vertrauen in Technologie. doi: 10.18419/OPUS-8677

Kortmann, B., Lunkenheimer, K., & Ehret, K. (Eds.). (2020). *eWAVE*.

Kortmann, B., & Schneider, E. W. (Eds.). (2004). *A handbook of varieties of English Volume 1: Phonology*. Berlin: De Gruyter Mouton.

Kotz, S., Ravignani, A., & Fitch, W. (2018). The Evolution of Rhythm Processing. *Trends in Cognitive Sciences*, *22*(10), 896–910. doi: 10.1016/j.tics.2018.08.002

Krashen, S. (1982). *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon Press.

Krenn, B., Schreitter, S., & Neubarth, F. (2017). Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application. *AI & SOCIETY*, *32*(1), 65–77. doi: 10.1007/s00146-014-0569-0

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. doi: 10.1080/01621459.1952.10483441

Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing* (Vol. 1, pp. 125–128). doi: 10.1109/PACRIM.1993.407206

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Kuo, F.-Y., Ouyang, I., Aryal, S., & Lanchantin, P. (2019). Selection and Training Schemes for Improving TTS Voice Built on Found Data. In *Proc. Interspeech 2019* (pp. 1516–1520). doi: 10.21437/Interspeech.2019-2816

Kurpaska, M. (2010). *Chinese Language(s): A Look through the Prism of The Great Dictionary of Modern Chinese Dialects*. DE GRUYTER MOUTON. doi: 10.1515/9783110219159

Labov, W. (1963). The Social Motivation of a Sound Change. *Word*, *19*(3), 273–309. doi: 10.1080/00437956.1963.11659799

Labov, W. (1966). *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics. doi: 10.1017/CBO9780511618208

Labov, W. (1969). *A study of non-standard english.* ERIC.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1984). Field Methods of the Project on Linguistic Change and Variation. In J. Baugh & J. Scherzer (Eds.), *Language in Use: Readings and Sociolinguistics* (pp. 28–66). Englewood Cliffs: Prentice Hall.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, *2*(2), 205–254. doi: 10.1017/S0954394500000338

Labov, W. (1994). *Principles of linguistic change Vol. 1, Internal factors*. Oxford: Oxford : Blackwell.

Labov, W. (2001). *Principles of linguistic change Vol. 2, Social factors*. Oxford: Oxford : Blackwell.
Labov, W. (2010). *Principles of linguistic change Vol. 3, Cognitive and cultural factors*. Oxford: Oxford : Blackwell.
Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. doi: 10.1121/1.1908694
Ladefoged, P., & Johnson, K. (2014). *A course in phonetics* (7. ed ed.). Stamford, CT: Cengage Learning.
Lado, R. (1957). Language across cultures. *Ann Arbor: University of Michigan Press*.
Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. doi: 10.2307/2529310
Lawrence, J. S., Crocker, J., & Blanton, H. (2011). Stigmatized and Dominant Cultural Groups Differentially Interpret Positive Feedback. *Journal of Cross-Cultural Psychology*, *42*(1), 165–169. doi: 10.1177/0022022110383569
Lee, H.-Y., & Song, J. (2019). Evaluating Korean learners' English rhythm proficiency with measures of sentence stress. *Applied Psycholinguistics*, *40*(6), 1363–1376. doi: 10.1017/S0142716419000298
Lee, M. W. (1987). Languages in China. In S. A. Wurm, R. Li, T. Baumann, & M. W. Lee (Eds.), *The Language Atlas of China*.
Lee, W.-S., & Zee, E. (2003). Illustrations of the IPA. Standard Chinese (Beijing). *Journal of the International Phonetic Association*, *33*(1), 109–112.
Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge University Press.
Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *5*, 253–264.
Leroux, W., & Kendall, T. (2018). English article acquisition by Chinese learners of English: An analysis of two corpora. *System*, *76*, 13–24.
Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*(6), 1093–1096. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022103110001459 doi: 10.1016/j.jesp.2010.05.025
Levene, H. (1960). Robust tests for equality of variances. In I. Olkin & H. Hotelling (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278–292). Stanford University Press.
Li, A., & Post, B. (2014). L2 Acquisition of Prosodic Properties of Speech Rhythm: Evidence from L1 Mandarin and German Learners of English. *Studies in Second Language Acquisition*, *36*(2), 223–255. doi: 10.1017/S0272263113000752
Li, A., Yu, J., Chen, J., & Wang, X. (2004). A contrastive study of standard chinese and shanghai-accented standard chinese. *From traditional phonology to modern speech processing, Festschrift for professor Wu Zongji's 95th birthday*, 253–288.
Li, F.-K. (1973). Languages and dialects of china. *Journal of Chinese Linguistics*, *1*(1), 1–13.
Li, S., & Sewell, A. (2012). Phonological Features of China English. *Asian Englishes*, *15*(2), 80–101.
Li, W. (1993). Zhongguo yingyu yu zhongguoshi yingyu (China English and Chinglish). *Waiyu Jiaoxue Yu Yanjiu (Foreign Language Teaching and Researching)*, *4*, 18–24.
Li, X., & Bisang, W. (2012). Classifiers in Sinitic languages: From individuation to definiteness-marking. *Lingua*, *122*(4), 335–355.
Liang, J. (2015). *Collocational features of China English: A corpus-based contrastive study* (Unpublished doctoral dissertation). The Hong Kong Institute of Education.

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.

Lim, L., & Ansaldo, U. (2020). Colloquial Singapore English (Singlish). In B. Kortmann, K. Lunkenheimer, & K. Ehret (Eds.), *The Electronic World Atlas of Varieties of English.*

Lin, C., & Jackson, L. (2021). Assimilation over protection: Rethinking mandarin language assimilation in China. *Multicultural Education Review*, *13*(4), 338–361. doi: 10.1080/2005615X.2021.2006117

Lin, L., Ginns, P., Wang, T., & Zhang, P. (2020). Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? *Computers & Education*, *143*, 103658. doi: 10.1016/j.compedu.2019.103658

Liu, J., & Ren, W. (2016). Apologies in Emails: Interactions Between Chinese EFL Learners and Their Foreign Peers. In Y.-s. Chen, D.-H. V. Rau, & G. Rau (Eds.), *Email Discourse Among Chinese Using English as a Lingua Franca* (pp. 205–228). Springer.

Liu, S., & Takeda, K. (2021). Mora-timed, stress-timed, and syllable-timed rhythm classes: Clues in English speech production by bilingual speakers. *Acta Linguistica Academica*. doi: 10.1556/2062.2021.00469

Liu, Y., Fang, A. C., & Wei, N. (2017). A Corpus-Based Study of Syntactic Patterns of Nominalizations Across Chinese and British Media English. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching Chinese English: The State of the Art* (pp. 77–92). Springer.

Llamas, C. (2007). Age. In C. Llamas, L. Mullany, & P. Stockwell (Eds.), *The Routledge Companion to Sociolinguistics* (??th ed.). Routledge. doi: 10.4324/9780203441497

Lloyd James, A. (1935). *The broadcast word*. London: Kegan Paul, Trench, Trubner & Co.

Lloyd James, A. (1940). *Speech signals in telephony*. London: Pitman.

Loakes, D., Hajek, J., & Fletcher, J. (2017). Can you t[æ]ll I'm from M[æ]lbourne?: An overview of the dress and trap vowels before /l/ as a regional accent marker in Australian English. *English World-Wide. A Journal of Varieties of English*, *38*(1), 29–49. doi: 10.1075/eww.38.1.03loa

Lobanov, B. M. (1971). 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

Lorenzoni, A., Pagliarini, E., Vespignani, F., & Navarrete, E. (2022). Pragmatic and knowledge range lenience towards foreigners. *Acta Psychologica*, *226*, 103572. doi: 10.1016/j.actpsy.2022.103572

Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America*, *129*(5), 3258–3270. doi: 10.1121/1.3559709

Louwerse, M. M., Graesser, A. C., McNamara, D. S., & Lu, S. (2009). Embodied conversational agents as conversational partners. *Applied Cognitive Psychology*, *23*, 1244–1255.

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. doi: 10.1177/00238309000430040301

Luelsdorf, P. A. (1983). On praguian functionalism and some extensions. In J. Vachek & L. Dušková (Eds.), *Praguiana: Some basic and less known aspects of the prague linguistic school* (Vol. 12, p. xi-xxx). John Benjamins.

Luke, K. K. (2000). Phonological re-interpretation: The assignment of Cantonese tones to English words. In *International conference on chinese linguistics.*

Ma, Q., & Xu, Z. (2017). The nativization of english in china. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching chinese english: The state of the art* (pp. 189–201). Springer.

MacKenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, *6*(s1), 20180061. doi: 10.1515/lingvan-2018-0061

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, *37*(7), 30–40. doi: 10.1145/176789.176792

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The Effects of Nonnative Accents on Listening Comprehension: Implications for ESL Assessment. *TESOL Quarterly*, *36*(2), 173–190. doi: 10.2307/3588329

Malisz, Z., O'Dell, M. L., Nieminen, T., & Wagner, P. (2017). Perspectives on Speech Timing: Coupled Oscillator Modeling of Polish and Finnish. *Phonetica*, *73*(3-4), 229–255. doi: 10.1159/000450829

Mandl, S., Brade, J., Bretschneider, M., Skripcak, A., Asbrock, F., Meyer, B., ... Strobel, A. (2023). Social perception of Embodied Digital Technologies interacting with humans. In *Intelligent Human Systems Integration (IHSI 2023) Integrating People and Intelligent Systems.* doi: 10.54941/ahfe1002836

Manikutty, S., Anuradha, N., & Hansen, K. (2007). Does culture influence learning styles in higher education? *International Journal of Learning and Change*, *2*(1), 70. Retrieved 2022-02-14, from `http://www.inderscience.com/link.php?id=14896` doi: 10.1504/IJLC.2007.014896

Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., ... Yu, Z. (2022). Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, *71*, 101255. doi: 10.1016/j.csl.2021.101255

Martha, A. S. D., & Santoso, H. (2019). The Design and Impact of the Pedagogical Agent: A Systematic Literature Review. *The Journal of Educators Online*, *16*(1). doi: 10.9743/jeo.2019.16.1.8

Marton, W. (1981). Contrastive analysis in the classroom. In J. Fisiak (Ed.), *Contrastive linguistics and the language teacher.* Oxford: Pergamon Press.

Mayer, R. E. (2014a). Introduction to Multimedia Learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed., pp. 1–24). Cambridge University Press. doi: 10.1017/CBO9781139547369.002

Mayer, R. E. (2014b). Principles Based on Social Cues in Multimedia Learning: Personalization, Voice, Image, and Embodiment Principles. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed., pp. 345–368). Cambridge University Press. doi: 10.1017/CBO9781139547369.017

Mayer, R. E., Dow, G. T., & Mayer, S. (2003). Multimedia Learning in an Interactive Self-Explaining Environment: What Works in the Design of Agent-Based Microworlds? *Journal of Educational Psychology*, *95*(4), 806–812. doi: 10.1037/0022-0663.95.4.806

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, *95*(2), 419–425. doi: 10.1037/0022-0663.95.2.419

McArthur, T. (1987). The English languages? *English Today*, *3*(3), 9–13. doi: 10.1017/S0266078400013511

McArthur, T. (1998). *The English Languages:* (1st ed.). Cambridge University Press. doi: 10.1017/9780511621048

McArthur, T. (2002). *The Oxford Guide to World English*. Oxford: Oxford University Press.

McAuliffe, M., Babel, M., & Vaughn, C. (2016). Do Listeners Learn Better from Natural Speech? In *Interspeech 2016* (pp. 2806–2810). ISCA. doi: 10.21437/Interspeech.2016-610

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

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017b). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017* (pp. 498–502). ISCA. doi: 10.21437/Interspeech.2017-1386

McAuliffe, M., & Sonderegger, M. (2022a). *English (nigeria) mfa dictionary v2.0.0* (Tech. Rep.). `https://mfa-models.readthedocs.io/pronunciationdictionary/English/English(Nigeria)MFAdictionaryv2_0_0.html`.

McAuliffe, M., & Sonderegger, M. (2022b). *English (nigeria) mfa g2p model v2.0.0* (Tech. Rep.). `https://mfa-models.readthedocs.io/G2Pmodel/English/English(Nigeria)MFAG2Pmodelv2_0_0.html`.

McAuliffe, M., & Sonderegger, M. (2022c). *English (uk) mfa dictionary v2.0.0* (Tech. Rep.). `https://mfa-models.readthedocs.io/pronunciationdictionary/English/English(UK)MFAdictionaryv2_0_0.html`.

McAuliffe, M., & Sonderegger, M. (2022d). *English (us) arpa acoustic model v2.0.0* (Tech. Rep.). `https://mfa-models.readthedocs.io/acoustic/English/English(US)ARPAacousticmodelv2_0_0.html`.

McAuliffe, M., & Sonderegger, M. (2022e). *English (us) arpa g2p model v2.0.0* (Tech. Rep.). `https://mfa-models.readthedocs.io/G2Pmodel/English/English(US)ARPAG2Pmodelv2_0_0.html`.

McCawley, J. D. (1992). Justifying Part-of-Speech Assignments in Mandarin Chinese. *Journal of Chinese Linguistics*, *20*(2), 211–246.

McCloy, D. R. (2016). *phonR: Tools for phoneticians and phonologists* [Manual].

McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communications Monographs*, *66*(1), 90–103. doi: 10.1080/03637759909376464

McKenzie, R. M., Kitikanan, P., & Boriboon, P. (2016). The competence and warmth of Thai students' attitudes towards varieties of English: The effect of gender and perceptions of L1 diversity. *Journal of Multilingual and Multicultural Development*, *37*(6), 536–550. doi: 10.1080/01434632.2015.1083573

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 101–116).

Mele, A. R. (2009). Intention and Intentional Action. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind* (1st ed., pp. 691–710). Oxford University Press. doi: 10.1093/oxfordhb/9780199262618.003.0041

Mesthrie, R., & Bhatt, R. M. (2008). *World Englishes: The Study of New Linguistic Varieties* (1st ed.). Cambridge University Press. doi: 10.1017/CBO9780511791321

Mesthrie, R., Kortmann, B., & Schneider, E. W. (Eds.). (2008). *Africa, South and Southeast Asia*. Mouton de Gruyter. doi: 10.1515/9783110208429

Method for the subjective assessment of intermediate quality level of audio systems. (2015). International Telecommunication Union Radiocommunication Assembly.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, *59*, 210–220. doi: 10.1016/j.pragma

.2013.07.012
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2022). *E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU wien* [Manual].
Meyer, S., Albrecht, S., Eibl, M., Rey, G. D., Schmied, J., Tamboli, R., ... Gesmann-Nuissl, D. (2022). Untying the Gordian Knot: Legally Compliant Sound Data Collection and Processing for TTS Systems in China. In *Privacy Symposium.* Venice: Springer.
Meyer, S., Mandl, S., Gesmann-Nuissl, D., & Strobel, A. (2022). Responsibility in Hybrid Societies: Concepts and terms. *AI and Ethics*, *3*(1), 25–48. doi: 10.1007/s43681-022-00184-2
Meyerhoff, M. (2011). *Introducing sociolinguistics* (2. ed ed.). London: Routledge.
Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(2), 339–384. doi: 10.1017/S0022226700010306
Milroy, L. (1997). Field Linguistics. In N. Coupland & A. Jaworski (Eds.), *Sociolinguistics: A Reader and Coursebook.* Basingstoke: Macmillan Press.
Milroy, L., & Gordon, M. J. (2003). *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.
Mok, P., & Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. doi: 10.5167/UZH-111798
Molenberghs, P. (2013). The neuroscience of in-group bias. *Neuroscience & Biobehavioral Reviews*, *37*(8), 1530–1536. doi: 10.1016/j.neubiorev.2013.06.002
Moran, D. (1996). A Case for Philosophical Pluralism: The Problem of Intentionality. *Royal Institute of Philosophy Supplement*, *40*, 19–32. doi: 10.1017/S135824610000583X
Moran, D. (2000). Heidegger?s Critique of Husserl?s and Brentano?s Accounts of Intentionality. *Inquiry*, *43*(1), 39–65. doi: 10.1080/002017400321361
Moreno, K. N., Klettke, B., Nibbaragandla, K., & Graesser, A. C. (2002). Perceived Characteristics and Pedagogical Efficacy of Animated Conversational Agents. In G. Goos, J. Hartmanis, J. Van Leeuwen, S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems* (Vol. 2363, pp. 963–971). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/3-540-47987-2_96
Moreno, R. (2012). Multimedia Learning with Animated Pedagogical Agents. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (1st ed., pp. 507–524). Cambridge University Press. doi: 10.1017/CBO9780511816819.032
Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, *83*(5), 405–408. doi: 10.1037/0033-295X.83.5.405
Murray, S. O. (1994). *Theory Groups and the Study of Language in North America: A Social History.* Amsterdam/Philadelphia: John Benjamins.
Myers, S. A., & Martin, M. M. (2018). Instructor credibility. In A. M. Hosek & M. L. Houser (Eds.), *Handbook of instructional communication: Rhetorical and relational perspectives* (pp. 38–50). Routledge.
Nass, C., & Lee, K. M. (2001a). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171–181. doi: 10.1037/1076-898X.7.3.171
Nass, C., & Lee, K. M. (2001b). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171–181. doi: 10.1037/1076-898X.7.3.171

Nearey, T. M. (1978). *Phonetic feature systems for vowels.* Indiana: Indiana University Linguistics Club.
Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113. doi: 10.1121/1.397861
Needham, J. (2005). *Science and civilisation in China. Vol. 2: History of scientific thought / by Joseph Needham with the research assistance of Wang Ling* (Reprint ed., Vol. 2). Cambridge: Cambridge Univ. Press.
Nelson, C. L., Proshina, Z. G., Davis, D. R., Kachru, B. B., & Kachru, Y. (Eds.). (2019). *Handbook of world Englishes* (2nd edition ed.). Hoboken, NJ: Wiley-Blackwell.
Nolan, F., & Asu, E. L. (2009). The Pairwise Variability Index and Coexisting Rhythms in Language. *Phonetica*, *66* (1-2), 64–77. doi: 10.1159/000208931
Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130396. doi: 10.1098/rstb.2013.0396
Nomura, J., & Ishikawa, K. (2018). Effects of first language processes and representations on second language perception: The case of vowel epenthesis by Japanese speakers. *International Journal of Bilingualism*, *22*(1), 69–87.
Norman, J. (1988). *Chinese*. Cambridge University Press.
NVIDIA. (2019). *WaveGlow LJS 256 channels.* https://catalog.ngc.nvidia.com/orgs/nvidia/models/waveglow-ljs-256channels.
O'Dell, M. L., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In *Proceedings of the XIVth international congress of phonetic sciences* (Vol. 2, pp. 1075–1078). University of California Berkeley.
O'Dell, M. L., & Nieminen, T. (2009). Coupled oscillator model for speech timing: Overview and examples. In *Nordic prosody: Proceedings of the 10th conference* (pp. 179–190). Helsinki.
Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America*, *54*(5), 1235–1247. doi: 10.1121/1.1914393
Ordin, M., & Polyanskaya, L. (2014). Development of timing patterns in first and second languages. *System*, *42*, 244–257. doi: 10.1016/j.system.2013.12.004
Ordin, M., & Polyanskaya, L. (2015a). Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. *The Journal of the Acoustical Society of America*, *138*(2), 533–544. doi: 10.1121/1.4923359
Ordin, M., & Polyanskaya, L. (2015b). Perception of speech rhythm in second language: The case of rhythmically similar L1 and L2. *Frontiers in Psychology*, *6* . doi: 10.3389/fpsyg.2015.00316
O'Shaughnessy, D. (1987). *Speech communication: Human and machine*. Reading, Mass: Addison-Wesley Pub. Co.
Paillereau, N., & Chládková, K. (2019). Spectral and temporal characteristics of Czech vowels in spontaneous speech. *AUC PHILOLOGICA*, *2019*(2), 77–95.
Palková, Z. (1994). *Fonetika a fonologie Češtiny: s obecnỳm úvodem do problematiky oboru [Phonetics and phonology of Czech: A general introduction to the field]*. Karolinum.
Park, K., & Kim, J. (2019). *g2pE.* `https://github.com/Kyubyong/g2p`. GitHub.
Paul, W. (2005). Adjectival modification in Mandarin Chinese and related issues. *Linguistics*, *43*(4), 757–793.
Pebesma, E. (2018). Simple features for r: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. doi: 10.32614/RJ-2018-009
Pennycook, A. (2010). *Language as a Local Practice* (??th ed.). Routledge. doi: 10.4324/

9780203846223
Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. doi: 10.1121/1.1906875
Pfitzinger, H. R., & Tamashima, M. (2006). Comparing perceptual local speech rate of German and Japanese speech. In *Proc. of the 3rd int. Conf. on speech prosody* (Vol. 1, pp. 105–108). TUD Press Dresden.
Phillipson, R. (1992). *Linguistic Imperialism*. Oxford: Oxford University Press.
Phillipson, R. (2010). *Linguistic Imperialism Continued*. New York: Routledge.
Phillipson, R. (2018). Linguistic Imperialism. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–7). Wiley. doi: 10.1002/9781405198431.wbeal0718.pub2
Pike, K. L. (1945). *The Intonation of American English*. Ann Arbor: University of Michigan Press.
Ping, W., Peng, K., Gibiansky, A., Arık, S. Ö., Kannan, A., Narang, S., ... Miller, J. (2018). Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. In *International conference on learning representations 2018.*
Podesva, R. J., & Zsiga, E. (2013). Sound recordings: Acoustic and articulatory data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 169–194). Cambridge University Press.
Podlipský, V. J., Šimáčková, Š., & Petráž, D. (2016). Is there an interlanguage speech credibility benefit? *Topics in Linguistics*, *17*(1), 30–44. doi: 10.1515/topling-2016-0003
Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *The Journal of the Acoustical Society of America*, *53*(4), 1093–1101. doi: 10.1121/1.1913429
Polyanskaya, L., & Ordin, M. (2019). The effect of speech rhythm and speaking rate on assessment of pronunciation in a second language. *Applied Psycholinguistics*, *40*(03), 795–819. doi: 10.1017/S0142716419000067
Port, R. F. (2003). Meter and speech. *Journal of Phonetics*, *31*(3-4), 599–611. doi: 10.1016/j.wocn.2003.08.001
Porzuczek, A. (2012). Measuring Vowel Duration Variability in Native English Speakers and Polish Learners. *Research in Language*, *10*(2), 201–214. doi: 10.2478/v10015-011-0034-9
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)* (pp. 3617–3621). doi: 10.1109/ICASSP.2019.8683143
Prieto, P., Vanrell, M. D. M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*, *54*(6), 681–702. doi: 10.1016/j.specom.2011.12.001
Prinz, W. (2012). *Open minds: The social making of agency and intentionality*. MIT Press.
Prinz, W. (2013). Self in the mirror. *Consciousness and cognition*, *22*(3), 1105–1113.
Prinz, W. (2017). Modeling self on others: An import theory of subjectivity and selfhood. *Consciousness and Cognition*, *49*, 347–362.
Pronunciation model: Hong kong english. (n.d.). Retrieved 2021-04-13.
Pustejovsky, J., & Krishnaswamy, N. (2021). Embodied Human Computer Interaction. *KI - Künstliche Intelligenz*, *35*(3-4), 307–327. doi: 10.1007/s13218-021-00727-5
Qiu, C., & Liang, J. (2015). The effect of stress on vowel space in daxi hakka chinese. In *Sixteenth annual conference of the international speech communication association.*
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English*

language. London: Longman.
R Core Team. (2023). *R: A language and environment for statistical computing* [Manual]. Vienna, Austria.
Ramsey, S. R. (1989). *The Languages of China*. Princeton: Princeton University Press.
Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. In *Speech prosody 2002, international conference.*
Ramus, F., Dupoux, E., & Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In *15th international congress of phonetic sciences* (pp. 337–342). Barcelona, Spain.
Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292. doi: 10.1016/S0010-0277(99)00058-X
Ran, Y., van de Weijer, J., & Sloos, M. (2020). Intonation in Hong Kong English and Guangzhou Cantonese-accented English: A Phonetic Comparison. *Journal of Language Teaching and Research*, *11*(5), 724–738.
Rawassizadeh, R., Sen, T., Kim, S. J., Meurisch, C., Keshavarz, H., Mühlhäuser, M., & Pazzani, M. (2019). Manifestation of virtual assistants and robots into daily life: Vision and challenges. *CCF Transactions on Pervasive Computing and Interaction*, *1*(3), 163–174. doi: 10.1007/s42486-019-00014-1
Reichelt, M., Kämmerer, F., Niegemann, H. M., & Zander, S. (2014). Talk to me personally: Personalization of language style in computer-based learning. *Computers in Human Behavior*, *35*, 199–210. doi: 10.1016/j.chb.2014.03.005
Ren, W. (2017). Pragmatics in Chinese graduate students' English gratitude emails. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching chinese english: The state of the art* (pp. 109–124). Springer.
Reuter, M. (1999). Merleau-Ponty's Notion of Pre-Reflective Intentionality. *Synthese*, *118*(1), 69–88. doi: 10.1023/A:1005144911619
Rey, G. D., & Steib, N. (2013). The personalization effect in multimedia learning: The influence of dialect. *Computers in Human Behavior*, *29*(5), 2022–2028.
Rice, L. (1976). Hardware & software for speech synthesis. *Dr. Dobb's Journal of Computer Calisthenics & Orthodontia*, *1*(4), 6–8.
Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (Ed.), *Linguistic controversies* (pp. 73–79). London: Edward Arnold.
Robb, M., & Chen, Y. (2008). Vowel Space in Mandarin-Accented English. *Asia Pacific Journal of Speech, Language and Hearing*, *11*(3), 175–188. doi: 10.1179/136132808805297179
Rosenfelder, I., Fruehwald, J., Keelan Evanini, Seyfarth, S., Gorman, K., Prichard, H., & Jiahong Yuan. (2015). *FAVE: Speaker fix.* Zenodo. doi: 10.5281/ZENODO.22281
RStudio Team. (2023). *RStudio: Integrated development environment for R* [Manual]. Boston, MA.
Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43-49. doi: 10.1109/TASSP.1978.1163055
Sandygulova, A., & O'Hare, G. (2015). Children's perception of synthesized voice: Robot's gender, age, and accent. In A. Tapus, E. André, J.-C. Martin, F. Ferland, & M. Ammi (Eds.), *Social robotics. 7th international conference, ICSR 2015.* Springer International Publishing.
Sankoff, D., & Laberge, S. (1978). The Linguistic Market and the Statistical Explanation of Variability. In D. Sankoff (Ed.), *Linguistic Variation: Models and Methods* (pp. 239–250). New York: Academic Press.

Santa Ana, O. A. (1996). Sonority and syllable structure in Chicano English. *Language Variation and Change*, *8*(1), 63–89. doi: 10.1017/S0954394500001071
Sapir, E. (1929). The Status of Linguistics as a Science. *Language*, *5*(4), 207. doi: 10.2307/409588
Scharinger, M., Monahan, P. J., & Idsardi, W. J. (2011). You had me at "Hello": Rapid extraction of dialect information from spoken words. *NeuroImage*, *56*(4), 2329–2338. doi: 10.1016/j.neuroimage.2011.04.007
Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. In *Proc. of the ICPhS* (pp. 607–610).
Schilling-Estes, N. (1998). Investigating "self-conscious" speech: The performance register in Ocracoke English. *Language in Society*, *27*(1), 53–83. doi: 10.1017/S0047404500019722
Schlüter, J. (2005). *Rhythmic Grammar: The Influence of Rhythm on Grammatical Variation and Change in English*. Mouton de Gruyter. doi: 10.1515/9783110219265
Schmied, J. (1991a). *English in Africa. An Introduction.* London: Longman.
Schmied, J. (1991b). National and subnational features in Kenyan English. In J. Cheshire (Ed.), *English around the World: Sociolinguistic perspectives* (pp. 420–32). Cambridge University Press Cambridge.
Schmied, J. (2020). Limits of Discourse: Examples from Political, Academic, and Human-Agent Interaction. *Discourse and Interaction*, *13*(2), 89–118. doi: 10.5817/DI2020-2-89
Schneider, E. W. (2007). *Postcolonial english: Varieties around the world*. Cambridge University Press.
Schneider, E. W. (2011). English into Asia: From Singaporean Ubiquity to Chinese Learners' Features. In M. Adams & A. Curzan (Eds.), *Contours of English and English Language Studies* (pp. 134–156). Michigan: University of Michigan Press.
Schneider, E. W. (2014a). *Models of English in the World* (Vol. 1; M. Filppula, J. Klemola, & D. Sharma, Eds.). Oxford University Press. doi: 10.1093/oxfordhb/9780199777716.013.001
Schneider, E. W. (2014b). New reflections on the evolutionary dynamics of world Englishes. *World Englishes*, *33*(1), 9–32.
Schneider, E. W., & Kirkpatrick, A. (2010). Developmental Patterns of English. In *The Routledge Handbook of World Englishes* (pp. 372–384). New York: Routledge.
Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2022). Psychologische Befunde zum Lernen mit digitalen Medien – ein Überblick. In M. A. Pfannstiel & P. F.-J. Steinhoff (Eds.), *E-Learning im digitalen Zeitalter* (pp. 581–605). Wiesbaden: Springer Fachmedien Wiesbaden. doi: 10.1007/978-3-658-36113-6_28
Schreier, D. (2005). *Consonant Change in English Worldwide: Synchrony meets Diachrony*. New York: Palgrave Macmillan.
Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How Effective are Pedagogical Agents for Learning? A Meta-Analytic Review. *Journal of Educational Computing Research*, *49*(1), 1–39. doi: 10.2190/EC.49.1.a
Schroeder, N. L., Chiou, E. K., & Craig, S. D. (2021). Trust influences perceptions of virtual humans, but not necessarily learning. *Computers & Education*, *160*, 104039. doi: 10.1016/j.compedu.2020.104039
Schroeder, N. L., & Gotch, C. M. (2015). Persisting Issues in Pedagogical Agent Research. *Journal of Educational Computing Research*, *53*(2), 183–204. doi: 10.1177/0735633115597625
Schuchardt, H. (1885). *über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Oppenheim.
Schultze-Berndt, E. (2006). Linguistic annotation. In J. Gippert, N. Himmelmann, & U. Mosel (Eds.),

Essentials of language documentation (pp. 213–225). Berlin: Mouton de Gruyter.
Schwartz, B. D., & Sprouse, R. (1994). Word Order and Nominative Case in Nonnative Language Acquisition: A Longitudinal Study of (l1 Turkish) German Interlanguage. In T. Hoekstra, B. D. Schwartz, H. Clahsen, & W. Rutherford (Eds.), *Language Acquisition Studies in Generative Grammar* (Vol. 8, pp. 317–368). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research*, *12*(1), 40–72.
Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2022). Voice in Human–Agent Interaction: A Survey. *ACM Computing Surveys*, *54*(4), 1–43. doi: 10.1145/3386867
Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. doi: 10.1017/S0140525X00005756
Searle, J. R. (1983). *Intentionality, an essay in the philosophy of mind*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.
Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge: MIT Press.
Sebastiangalles, N. (1992). Contrasting syllabic effects in Catalan and Spanish*1. *Journal of Memory and Language*, *31*(1), 18–32. doi: 10.1016/0749-596X(92)90003-G
Setter, J. (2003). A comparison of speech rhythm in British and Hong Kong English. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 467–470).
Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. doi: 10.1093/biomet/52.3-4.591
Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (p. 4779-4783). doi: 10.1109/ICASSP.2018.8461368
Siebenhaar, B., & Hahn, M. (2019). Vowel space, speech rate and language space. In *Proceedings of the 19th international congress of phonetic sciences. Melbourne, australia.*
Simo Bobda, A. (2008). Cameroon English: Phonology. In R. Mesthrie, B. Kortmann, & E. W. Schneider (Eds.), *Varieties of English 4 Africa, South and Southeast Asia.* Berlin: Mouton de Gruyter.
Snape, N., Leung, Y.-k. I., & Ting, H.-C. (2006). Comparing Chinese, Japanese and Spanish Speakers in L2 English Article Acquisition: Evidence against the Fluctuation Hypothesis? In *8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006)* (pp. 132–139).
Sotelo, J. M. R., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A. C., & Bengio, Y. (2017). Char2Wav: End-to-End Speech Synthesis. In *ICLR*.
Souza, A. L., & Markman, A. B. (2013). Foreign accent does not influence cognitive judgments. In *Proceedings of the annual meeting of the cognitive science society.*
Spolsky, B. (2010). Ferguson and Fishman: Sociolinguistics and the Sociology of Language. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics.* SAGE.
Sprouse, R. A., & Schwartz, B. D. (1998). In defense of full transfer in German-English and French-English interlanguage: Comparative L2 acquisition research. In *Proceedings of the Annual Boston University Conference on Language Development* (Vol. 22, pp. 726–736).
Stockmal, V., Markus, D., & Bond, D. (2005). Measures of Native and Non-Native Rhythm in a Quan-

tity Language. *Language and Speech*, *48*(1), 55–63. doi: 10.1177/00238309050480010301
Strang, B. M. H. (1970). *A history of English*. London: Meuthen.
Strange, W., Bohn, O.-S., Nishi, K., & Trent, S. A. (2005). Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *The Journal of the Acoustical Society of America*, *118*(3), 1751–1762. doi: 10.1121/1.1992688
Sun, L. (1989). Yingyu guobie bianti de yanjiu he yingyuzaizhongguo [Research on English varieties and English in China]. *Waiguo Yu [Foreign Languages]*, *60*(2), 17–23.
Sun, T. (2006). Zhongguo Yingyu de tezheng ji fazhan qianjing [The characteristics of China English and its future]. *Fujian nonglin daxue xuebao [Journal of Fujian Agriculture and Forestry University]*, *9*(1), 84–87.
Sutton, S. J., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In *CHI 2019*. Glasgow, Scotland, UK.
Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100. doi: 10.1121/1.393381
Tabain, M., Fletcher, J., & Butcher, A. (2014). Lexical stress in Pitjantjatjara. *Journal of Phonetics*, *42*, 52–66. doi: 10.1016/j.wocn.2013.11.005
Tabossi, P., Collina, S., Mazzetti, M., & Zoppello, M. (2000). Syllables in the processing of spoken Italian. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 758–775. doi: 10.1037/0096-1523.26.2.758
Tacotron 2 - PyTorch implementation with faster-than-realtime inference. (2018). `https://github.com/NVIDIA/tacotron2`. GitHub.
Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge University Press.
Tagliamonte, S. A. (2012). *Variationist sociolinguistics: Change, observation, interpretation*. Oxford: Wiley-Blackwell.
Tagliamonte, S. A. (2013). Analysing and Interpreting Variation in the Sociolinguistic Tradition. In M. Krug & J. Schlüter (Eds.), *Research Methods in Language Variation and Change* (pp. 382–402). Cambridge: Cambridge University Press.
Taiwo, R. (2020). Nigerian english. In B. Kortmann, K. Lunkenheimer, & K. Ehret (Eds.), *The electronic world atlas of varieties of english.*
Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., & Broadbent, E. (2011). The Effects of Synthesized Voice Accents on User Perceptions of Robots. *International Journal of Social Robotics*, *3*(3), 253–262. doi: 10.1007/s12369-011-0100-4
Tamboli, R. (in prep.). Establishing Conversational Pedagogical Agents as Credible Knowledge Providers: The Case of Synthsized Indian English.
Tang, C., & van Heuven, V. J. (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, *119*(5), 709–732. doi: 10.1016/j.lingua.2008.10.001
Taubert, S. (2022a). *Dict-from-g2pE.* Zenodo. doi: 10.5281/ZENODO.7129911
Taubert, S. (2022b). *Tacotron-cli.* Computer software.
Taubert, S. (2022c). *Waveglow-cli.* Computer software.
Taubert, S. (2023a). *Textgrid-tools.* Zenodo. doi: 10.5281/ZENODO.7986716
Taubert, S. (2023b). *Txt-utils.* Zenodo. doi: 10.5281/ZENODO.7986310
Taubert, S., Sternkopf, J., Kahl, S., & Eibl, M. (2022). A Comparison of Text Selection Algorithms

for Sequence-to-Sequence Neural TTS. In *2022 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)* (pp. 1–6). Xi'an, China: IEEE. doi: 10.1109/ICSPCC55723.2022.9984283

Tavani, H. (2018). Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information*, *9*(4), 73. doi: 10.3390/info9040073

Taylor, P. A. (2009a). *Text-to-speech synthesis*. Cambridge university press.

Taylor, P. A. (2009b). *Text-to-speech synthesis*. Cambridge university press.

Temperley, D. (2009). Distributional Stress Regularity: A Corpus Study. *Journal of Psycholinguistic Research*, *38*(1), 75–92. doi: 10.1007/s10936-008-9084-0

Teven, J. J. (2007). Teacher Caring and Classroom Behavior: Relationships with Student Affect and Perceptions of Teacher Competence and Trustworthiness. *Communication Quarterly*, *55*(4), 433–450. doi: 10.1080/01463370701658077

Thomas, E., R. (2002). Instrumental phonetics. In P. Trudgill, J. K. Chambers, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (pp. 168–200). Oxford: Blackwell.

Thomas, E., R., & Kendall, T. (2007). *NORM: The vowel normalization and plotting suite.*

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100. doi: 10.1121/1.399849

Trudgill, P. (2000). *Sociolinguistics: An Introduction to Language and Society* (4th ed ed.). London, New York: Penguin.

Trudgill, P. (2004). *New-dialect formation: The inevitability of colonial Englishes*. Edinburgh: Edinburgh university press.

Trudgill, P. (2008). Colonial dialect contact in the history of european languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, *37*(2), 241–254.

Trudgill, P., & Hannah, J. (2013). *International English* (??th ed.). Routledge. doi: 10.4324/9780203785225

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. doi: 10.1093/mind/LIX.236.433

Turk, A., & Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology*, *4*(1). doi: 10.1515/lp-2013-0005

Vandevelde, P. (2005). Derrida's Intentional Skepticism: A Husserlian Response. *Journal of the British Society for Phenomenology*, *36*(2), 160–178. doi: 10.1080/00071773.2005.11006540

Van Santen, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication*, *11*(6), 513–546. doi: 10.1016/0167-6393(92)90027-5

Van Santen, J. P. H., & Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *The Journal of the Acoustical Society of America*, *107*(2), 1012–1026. doi: 10.1121/1.428281

Venezky, R. L. (1995). How English is Read: Grapheme-Phoneme Regularity and Orthographic Structure in Word Recognition. In R. M. Joshi, I. Taylor, & D. R. Olson (Eds.), *Scripts and Literacy* (Vol. 7, pp. 111–129). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-011-1162-1_8

Volín, J. (2017). Appeal and disrepute of the so-called global rhythm metrics. *Acta Universitatis Carolinae Philologica*(3), 79–94.

Volín, J., & Skarnitzl, R. (2010). The strength of foreign accent in Czech English under adverse listening

conditions. *Speech Communication*, *52*(11-12), 1010–1021. doi: 10.1016/j.specom.2010.06.009

Wade, L. (2017). The role of duration in the perception of vowel merger. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 30. doi: 10.5334/labphon.54

Wang, H., & van Heuven, V. J. (2006). Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. *Linguistics in the Netherlands*, *23*, 237–248. doi: 10.1075/avt.23.23wan

Wang, H., & van Heuven, V. J. (2018). Relative Contribution of Vowel Quality and Duration to Native Language Identification in Foreign-Accented English. In *Proceedings of the 2nd International Conference on Cryptography, Security and Privacy* (pp. 16–20). Guiyang China: ACM. doi: 10.1145/3199478.3199507

Wang, M., & Koda, K. (2007). Commonalities and Differences in Word Identification Skills Among Learners of English as a Second Language: Commonalities and Differences in Word Identification Skills. *Language Learning*, *57*, 201–222. doi: 10.1111/j.1467-9922.2007.00416.x

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.

Wang, Q. (2023). An instrumental investigation of the vowel inventory of China English. *Asian Englishes*, *25*(1), 111–132. doi: 10.1080/13488678.2022.2158265

Wang, R. (1991). Zhongguo yingyu shi keguan cunzai [China English is an objective reality]. *Jiefangjun Waiyu Xueyuan Xuebao [Journal of PLA Foreign Languages Institute]*, *1*(1), 1–8.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. In *Interspeech 2017* (pp. 4006–4010). ISCA. doi: 10.21437/Interspeech.2017-1452

Wassink, A. B. (2006). A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties. *The Journal of the Acoustical Society of America*, *119*(4), 2334–2350. doi: 10.1121/1.2168414

Watt, D. (2000). Phonetic parallels between the close-mid vowels of Tyneside English: Are they internally or externally motivated? *Language Variation and Change*, *12*(1), 69–101. doi: 10.1017/S0954394500121040

Watt, D., & Fabricius, A. H. (2002). Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1-F2 plane. *Leeds Working Papers in Linguistics and Phonetics*, *9*, 159–173.

WaveGlow: a Flow-based Generative Network for Speech Synthesis. (2018). `https://github.com/NVIDIA/waveglow`. GitHub.

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

Weber, H., & Platt, J. T. (1980). *English in Singapore and Malaysia: Status, Features, Functions*. Oxford: Oxford University Press.

Wee, L.-H. (2008). Phonological Patterns in the Englishes of Singapore and Hong Kong. *World Englishes*, *27*(3-4), 480–501.

Wei, Y., & Fei, J. (2003). Using English in China. *English Today*, *19*(4), 42.

Weinreich, U. (1953). *Languages in Contact: Findings and Problems*. New York: Publications of the Linguistic Circle of New York 1.

Weinreich, U., Labov, W., & Herzog, M. (1968). Empirical Foundations for a Theory of Language Change. In W. P. Lehmann & Y. Malkiels (Eds.), *Directions for Historical Linguistics: A Sympo-*

sium. Austin: University of Texas Press.
Wells, J. C. (1982). *Accents of English: An Introduction* (Vol. 1). Cambridge: Cambridge University Press.
Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore, & R. Winski (Eds.), *Handbook of standards and resources for spoken language systems* (chap. Part 4, section B). Berlin / New York: Mouton de Gruyter.
Wen, Q. (2018). The production-oriented approach to teaching university students English in China. *Language Teaching*, *51*(4), 526–540. doi: 10.1017/S026144481600001X
White, K. (1985). The Meaning of Phantasia in Aristotle's De Anima , III, 3–8. *Dialogue*, *24*(3), 483–505. doi: 10.1017/S0012217300040348
White, L. (2002). *English speech timing: A domain and locus approach* (Unpublished doctoral dissertation). University of Edinburgh.
White, L. (2008). Different? Yes. Fundamentally? No. Definiteness Effects in the L2 English of Mandarin Speakers. In *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007)* (pp. 251–261).
White, L. (2014). Communicative function and prosodic form in speech timing. *Speech Communication*, *63–64*, 38–54. doi: 10.1016/j.specom.2014.04.003
White, L. (2018). Segmentation of speech. In S.-A. Rueschemeyer & G. Gaskell (Eds.), *Oxford handbook of psycholinguistics*. Oxford: Oxford University Press.
White, L., Luche, C. D., & Floccia, C. (2016). Five-month-old infants' discrimination of unfamiliar languages does not accord with "rhythm class". In *Speech Prosody 2016* (pp. 567–571). ISCA. doi: 10.21437/SpeechProsody.2016-116
White, L., & Malisz, Z. (2020). Speech rhythm and timing. In C. Gussenhoven & A. Chen (Eds.), *Oxford handbook of language prosody* (pp. 167–182). Oxford: Oxford University Press.
White, L., & Mattys, S. L. (2007a). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, *35*(4), 501–522. doi: 10.1016/j.wocn.2007.02.003
White, L., & Mattys, S. L. (2007b). Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), *Current Issues in Linguistic Theory* (Vol. 282, pp. 237–257). Amsterdam: John Benjamins Publishing Company. doi: 10.1075/cilt.282.16whi
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. doi: 10.1016/j.jml.2011.12.010
White, N. P. (1987). *Plato on knowledge and reality* (3. [unveränd.] Nachdr ed.). Indianapolis, Ind: Hackett.
Whiteside, S. P. (1998). Identification of a Speaker's Sex: A Study of Vowels. *Perceptual and Motor Skills*, *86*(2), 579–584. doi: 10.2466/pms.1998.86.2.579
Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686
Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation* [Manual].
Wiese, R. (1997). Underspecifícation and the description of Chinese vowels. In J. Wang & N. Smith (Eds.), *Studies in Chinese Phonology*. Berlin, Boston: DE GRUYTER. doi:

10.1515/9783110822014.219
Wiese, R. (2022). Rhythmic structure – parallels between language and music. In M. Scharinger & R. Wiese (Eds.), *How Language Speaks to Music* (pp. 135–160). De Gruyter. doi: 10.1515/9783110770186-006
Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *The Journal of the Acoustical Society of America*, *127*(3), 1559–1569. doi: 10.1121/1.3293004
Wodak, R., Johnstone, B., & Kerswill, P. (2010). Introduction. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics.* SAGE.
Wolfram, W. (2000). Issues in Reconstructing Earlier African-American English. *World Englishes*, *19*(1), 39–58. doi: 10.1111/1467-971X.00154
Wong, M. L.-Y. (2020). Hong Kong English. In B. Kortmann, K. Lunkenheimer, & K. Ehret (Eds.), *The Electronic World Atlas of Varieties of English.*
Xie, Z. (1995). Zhongguo yingyu: Kua wenhua jiaoji zhong de ganrao bianti (China English: Interference variety in cross-cultural communication). *Xiandai Waiyu [Modern Foreign Languages]*, *4*, 7–11.
Xu, Z. (2006). Rectifying "Chinese English". In A. Hashim & N. Hassan (Eds.), *Varieties of English in Southeast Asia and beyond* (pp. 283–291). Kuala Lumpur: University of Malaya Press.
Xu, Z. (2008). Analysis of Syntactic Features of Chinese English. *Asian Englishes*, *11*(2), 4–31.
Xu, Z. (2010). *Chinese English: Features and implications*. Hong Kong: Open University of Hong Kong Press.
Xu, Z. (2017). Researching chinese english: A meta-analysis of chinese scholarship on chinese english research. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching chinese english: The state of the art* (pp. 235–266). Springer.
Xu, Z. (2020). Chinese English. In A. Kirkpatrick (Ed.), *The Routledge Handbook of World Englishes* (pp. 265–297). New: Routledge.
Xu, Z., Deterding, D., & He, D. (2017). What we know about Chinese English: Status, issues and trends. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching chinese english: The state of the art* (pp. 1–14). Springer.
Xue, S. A., Hagstrom, F., & Hao, J. (2002). Speaking fundamental frequency characteristics of young and elderly bilingual Chinese-English speakers: A functional system approach. *Asia Pacific Journal of Speech, Language and Hearing*, *7*(1), 55–62.
Yan, M. M. (2006). *Introduction to Chinese dialectology* (No. 22). München: LINCOM Europa.
Yang, J. (2009). Chinese borrowings in English. *World Englishes*, *28*(1), 90–106.
Yazawa, K., Konishi, T., & Kondo, M. (2015). Vowel Epenthesis in Japanese Speakers' L2 English. In *ICPhS.*
Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the Brain to Weight Speech Cues Differently: A Study of Finnish Second-language Users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332. doi: 10.1162/jocn.2009.21272
Zee, E., & Lee, W.-S. (2007). Vowel Typology in Chinese. In *Proc. 16th ICPhS* (pp. 1429–1432).
Zhang, H. (2002). Bilingual creativity in Chinese English: Ha Jin's In the Pond. *World Englishes*, *21*(2), 305–315.
Zhang, X., & Ardasheva, Y. (2019). Sources of college EFL learners' self-efficacy in the English public

speaking domain. *English for Specific Purposes*, *53*, 47–59. doi: 10.1016/j.esp.2018.09.004
Zheng, Y., & Cheng, L. (2008). Test Review: College English Test (CET) in China. *Language Testing*, *25*(3), 408–417.
Zwitserlood, P., Schriefers, H., Lahiri, A., & Van Donselaar, W. (1993). The role of syllables in the perception of spoken Dutch. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 260–271. doi: 10.1037/0278-7393.19.2.260

A Deutsche Zusammenfassung der Dissertation

A.1 Einleitung

Die vorliegende kumulative Arbeit stellt eine Untersuchung nicht-muttersprachlicher Englischsprecher aus den chinesischen Provinzen Guangdong, Guangxi und Yunnan vor, die im Rahmen des Teilprojekts *Glaubwürdige pädagogische Konversationsagenten* des Sonderforschungsbereichs *Hybride Gesellschaften: Menschen interagieren mit verkörperten digitalen Technologien* angefertigt wurde. Der Sonderforschungsbereich befasst sich mit der Frage, wie Menschen und verkörperte digitale Technologien (z.B. humanoide Roboter, aber auch autonome Fahrzeuge oder digitale Assistenten) in einer zukünftigen hybriden Gesellschaft zusammenleben. Im Projekt, das dieser Dissertation zugrunde liegt, wird der Einfluss eines pädagogischen Konversationsagenten, der die gleiche nicht-muttersprachliche Varietät wie seine Benutzer spricht, auf Lernergebnisse untersucht. Das Projekt baut auf Vorarbeiten auf, die gezeigt haben, dass ein menschlicher Lehrer in Österreich, der anstatt der Standartvarietät die örtliche Varietät spricht, zu höheren Lernergebnissen der Schüler führt (Rey & Steib, 2013). Dieses Prinzip wird nun zum einen in den Fremdsprachkontext und zum anderen auf einen nicht menschlichen Lehrer (einen pädagogischen Konversationsagenten) übertragen. Die Persönlichkeit und Verkörperung des Agenten wird dazu (vorerst) nur über die Stimme realisiert.

Um diese übergeordnete Forschungsfrage zu beantworten, wird im ersten Schritt die bestehende Literatur zum chinesischen Englisch gesichtet und eine Systematik erstellt (Kapitel 2). Im Anschluss (Kapitel 3) wird der Vokalraum von Sprechern aus den Provinzen Guangdong, Guangxi und Yunnan mit Methoden der quantitativen Variationslinguistik beschrieben und auf mögliche *mergers*, also Vokalpaare die, anders als in anderen Varietäten wie z.B. dem muttersprachlichen Englisch, nicht differenziert werden, untersucht. Im nächsten Schritt (Kapitel 4) wird die Frage untersucht, ob chinesisches Englisch eine silbenzählende (*syllable-timed*) oder eine akzentzählende (*stress-timed*) Varietät ist. Dabei wird auch ein Vergleich mit anderen nicht-muttersprachlichen Varietäten gezogen. Danach (Kapitel 5) wird die, in der Analyse der Vokalräume verwendete Methodologie ausgebaut und eine darauf basierende, linguistisch motivierte Qualitätsmetrik für neurale Sprachsynthesesysteme vorgestellt. Im letzten Schritt (Kapitel 6) wird der Bogen zur übergeordneten Forschungsfrage gespannt und eine Studie vorgestellt, in welcher untersucht wurde, ob ein nicht-muttersprachlicher pädagogischer Konversationsagent tatsächlich einen Einfluss auf die Lernergebnisse von Studierenden hat.

Die Innovation des Forschungsprojekts besteht darin, bestehende Forschung zur Glaubwürdigkeit von Robotern (Finkel & Krämer, 2022), virtuellen Lehrern (K. N. Moreno et al., 2002) und der Glaubwürdigkeit menschlicher Lehrer (Finn et al., 2009a) zu kombinieren und die Idee, das ein glaubwürdiger Lehrer den Lernerfolg von Studierenden verbessert, auf einen pädagogischen Konversationsagenten im Kontext einer möglichen zukünftigen hybriden Gesellschaft ausgedehnt. Die vorliegenden Arbeit is innovativ, sowohl auf der fachlichen als auch auf der methodischen Ebene.

Auf der fachlichen Ebene schlägt Kapitel 2 eine Systematik der sprachlichen Merkmale des chinesischen Englisch, ähnlich bereits etablierter nicht-muttersprachlicher Varietäten aus der Region, wie dem Hong Kong Englischen oder dem Singapur Englischen vor. Kapitel 3 fügt zum einen eine neue Analyse

stratifizierter Daten für zwei bereits untersuchte Provinzen (Guangxi und Yunnan), zu denen aber jeweils nur die Arbeiten einer Forschungsgruppe verfügbar sind, und zum anderen eine Beschreibung und Analyse einer bisher noch nicht untersuchten Provinz (Guangdong) hinzu. In Kapitel 4 wird eine vergleichende Einordnung muttersprachlicher und nicht-muttersprachlicher Varietäten des Englischen auf dem Sprachrhythmuskontinuum vorgestellt. Diese Studie führt erstmals chinesisches Englisch von Sprechern aus verschiedenen Provinzen explizit auf und verortet zusätzlich erstmals tschechisches Englisch und italienisches Englisch als eher silbenzählend bzw. akzentzählend auf dem Kontinuum. Eine weiterführende Beschreibung der Vokalräume der beiden Varietäten wurden in einem anderen Rahmen veröffentlicht (Albrecht & Schmied, 2023). Kapitel 6 schließlich präsentiert die Ergebnisse der ersten Studie zur Frage, ob ein nicht-muttersprachlicher Agent die Lernergebnisse Studierender verbessern kann.

Auf der methodischen Ebene wird in Kapitel 5 eine neuartige Qualitätsmetrik für Sprachsynthesesysteme vorgestellt, die sich auf die Qualität der generierten Vokale konzentriert und so einerseits ein Maß für die Natürlichkeit der generierten Sprach bildet, andererseits aber auch als Indikator für zusätzliches Training des Systems dienen kann. Weiterhin werden in allen vorgestellten Analysen konsequent quantitative, automatisierte Analysemethoden eingesetzt, sowohl zur automatischen Transkription von Audiodaten, also auch zur automatisierten Erstellung phonetischer Messungen und der anschließenden Auswertung. Die beschriebenen Methoden lassen sich auf Grund der ebenfalls veröffentlichten Analyseskripte (z.B. Albrecht, 2023b, 2023c) leicht auf andere Datensätze anwenden und entwickeln so quantitative Methoden im *World Englishes* Paradigma weiter.

Zusammengefasst lässt sich feststellen, dass die vorliegende Arbeit sowohl breiter gefasste Forschungsfragen beantwortet, wie beispielsweise den Effekt eines nicht-muttersprachlichen Agenten auf Lernerfolge in hybriden Gesellschaften, als auch sprachliche Details, wie den Vokalraum und Sprachrhythmus des chinesischen Englisch von Sprechern aus Guangdong, Guangxi und Yunnan, untersucht.

A.2 Zentrale Begriffe

Durch die zunehmende Digitalisierung der Gesellschaft (Hilbert, 2020) ergeben sich immer häufiger Situationen in denen Menschen mit verkörperten digitalen Technologien (*embodied digital technologies, EDTs* Mandl et al., 2023), sei es in Form von Robotern, autonomen Fahrzeugen oder digitalen Assistenten, interargieren. Um diese Interaktionen in zukünftigen hybriden Gesellschaften möglichst reibungslos zu gestallten, ist es nötig, dass EDTs als glaubwürdige Kommunikationspartner auftreten.

Diese Glaubwürdigkeit leitet sich aus dem Prinzip der Intentionalität ab, einem zentralen Konzept der Philosophie. welches besagt dass Geisteszustände immer direktional sind, sich also immer auf etwas beziehen. Obwohl diese Intentionalität im menschlichen Gehirn das Resultat biologischer Prozesse ist (Searle, 1992), scheint sie eine exklusiv menschliche Funktion höherer Ordnung des Gehirns zu sein (Searle, 1983). Ein spannender Beitrag zur Diskussion der Intentionalität (und schlussendlich des Bewusstseins) stammt von Prinz (2012), der die Idee der sozialen Spiegelung (*social mirroring*) postuliert, die besagt, dass wir Menschen unser Bewusstsein durch Spiegelung des Verhaltens anderer konstruieren. In Kombination mit Anthropomorphisierungsprozessen im Rahmen derer Menschen EDTs Intentionalität zuschreiben, können diese EDTs somit als relativ vollwertige Akteure in sozialen Spiegelungsprozessen auftreten. Im linguistischen Sinn wird Intentionalität häufig im Sinne der Intention gegenüber einem Kommunikationspartner gesehen und ist damit direkt mit Glaubwürdigkeit

verbunden (McCroskey & Teven, 1999). Linguistische Glaubwürdigkeit wird durch sprachliche Eigenschaften realisiert und drückt den guten Willen (*goodwill*) dem Kommunikationspartner gegenüber aus.

Die sprachliche Realisierung lässt sich hervoragend mit den Methoden der Variationslinguistik untersuchen, welche insbesondere durch die Arbeiten von Labov (1963, 1966, 1969) in den 1960er Jahren in den USA in der Soziolinguistik etabliertert wurden. Die Variationslinguistik legt großen Wert auf die Untersuchung der Vernakularsprache, also die Sprache in Kontexten in denen Menschen ihre Sprache am wenigstens kontrollieren, auf Verantwortung im Forschungsprozess, in diesem Fall die Analyse aller Variation und nicht nur der interessanten, und auf die Begrenzung des Kontexts der untersuchten Variable, also die Identifikation aller funktional equivalenten Varianten einer soziolinguistischen Variable (Tagliamonte, 2012). In den Jahrzehnten seit den grundlegenden Arbeiten Labovs wurden außer sozialer Klasse (z.B. Labov, 1966) auch Geschlecht, Alter, Ausdrucksweise, Sprachebene, Ethnie und Kultur untersucht (einen Guten Überblick zur Forschung in diesen Bereichen ermöglichen Chambers, 2009; Tagliamonte, 2013; Trudgill, 2000).

Die Forschungsmethoden der Variationslinguistik, die sich anfangs hauptsächlich mit muttersprachlicher Variation im Englischen befassten, wurden in den 1980er Jahren durch das Konzept der *World Englishes* (Kachru, 1984) auf nicht-muttersprachliche Varietäten ausgedehnt. Im Zuge der Dekolonialisierungsprozesse auf dem Gebiet des ehemaligen Britischen Weltreichs setzten Bestrebungen ein, den akademischen Diskurs über das in ehemaligen Kolonien gesprochene Englisch von der Norm des Muttersprachlers zu lösen und diese Sprachen als eigentständige Varietäten zu betrachten. Um die Vielzahl der so zu betrachtenden Varietäten zu systematisieren wurden verschiedene Klassifikationsmodelle vorgeschlagen. Die beiden verbreitetsten Modelle sind das Modell der konzentrischen Kreise (Kachru, 1985) und das dynamische Modell (E. W. Schneider, 2007). Beide Modelle klassifizieren Varietäten je nach Status und Verwendung der englischen Sprach auf dem jeweiligen Gebiet.

Nach Kachrus Modell ist chinesisches Englisch eine Varietät des äußersten Kreises, in der noch sehr stark der Einfluss lokaler Sprachen sichtbar ist und dieser noch nicht abschließend kodifiziert ist. Daher sollte die vielfältige Sprachenlandschaft der VR China bei den Diskussionen des chinesischen Englischs beachtet werden. In China werden eine Vielzahl unterschiedlicher Sprachen, mit einer gemeinsamen Schrift, die aber nicht mit der Aussprache korreliert, gesprochen, die zu mindestens sechs Sprachfamilien gehören. Über die Klassifizierung der Sprachen herrscht in der Wissenschaft kein Konsens, jedoch darüber, dass sie in zwei große Gruppen eingeteilt werden können: die nördliche Mandarin Gruppe und die südliche Gruppe. Die Sprachen unterscheiden sich stark voneinander, sodass Sprecher einzelner Sprachen sich gegenseitig oft nicht verstehen. Sowohl die Vokalinventare als auch die Silbenstruktur und die lexikalischen Töne unterscheiden sich signifikant voneinander.

Diese zentralen Begriffe bilden den Forschungsrahmen der in dieser Dissertation vorgestellten Studien und sollten bei der Interpretation und Diskussion der Ergebnisse bedacht werden.

A.3 Ergebnisse

Die in Kapitel 3 präsentierte Analyse des Vokalraums des chinesischen Englisch, das in Guangdong, Guangxi und Yunnan gesprochen wird, konnte einige der impressionistischen Beschreibungen aus der

Literatur bestätigen. So zeigte sich, dass die gespannten (*tense*) und ungespannten[1] (*lax*) Vokalräume ähnlich groß sind und damit weniger verschieden voneinander sind als im muttersprachlichen Englisch. Weiterhin ist in den hier analysierten Daten der gerundete halboffene Hinterzungenvokal (CLOTH) recht weit hinten im Vergleich mit früheren Beschreibungen des chinesisch Englischen Vokalraums, jedoch ähnlich wie in Q. Wang (2023). Im Allgemeinen zeigt die Analyse, dass die gespannten und ungespannten Vokale qualitativ wenig unterschiedlich sind und sich nicht klar voneinander abgrenzen lassen, zumal das Vokalinventar des chinesischen Englisch quantitativ hauptsächlich volle Vokale enthält. Die Vokalpaare FLEECE-KIT und GOOSE-FOOT werden in den analysierten Daten nicht unterschieden, das Vokalpaar TRAP-DRESS hingegen wird unterschieden. Basierend auf diesen Ergebnissen wurde ein vorläufiges Vokalinventar des chinesichen English vorgeschlagen.

Die Analyse des Variabilität der Intervalldauer von Vokalen und Konsonanten verschiedener muttersprachlicher und nicht-muttersprachlicher Varietäten zeigte quantitativ, dass das chinesische Englisch in der Tat eine sich klar von muttersprachlichen Varietäten unterschiedliche Variabilität zeigt. Das tschechische Englisch und das italienische Englisch haben ebenfalls von den muttersprachlichen Varietäten abweichende Variabilität, die auf einen Einfluss der Muttersprache hindeutet. Die Analyse hat hat gezeigt, dass die eingesetzten Metriken sich eigenen, den Einfluss einer Muttersprache auf die Fremdsprache zu messen.

Die verwendeten Methoden haben gezeigt, dass automatische Transkription und Analyse valide Methoden der quantitativen Variationslinguistik sind. Für die automatische Transkription gilt es jedoch zu bedenken, dass eine gewisse grundlegende Qualität der Daten nötig ist. Insbesondere Daten weniger sicherer Sprecher, die oft mehrfach ansetzen ein Wort auszusprechen, führt zu fehlerhaften Transkriptionen. Die auf den eingesetzten Methoden basierte Qualitätsmetrik für Sprachsynthesesysteme unterstreicht die Validität der Methoden und erlaubt Einblicke in die Qualität synthetisierter Vokale die eine weitere Optimierung des Synthesesystems ermöglichen. Weiterhin hat die Analyse in Kapitel 3 gezeigt dass die Art der Datenerhebung, also ob Probanden einen Text vorlesen oder Interviewfragen beantworten, keinen signifikanten Einfluss auf den Vokalraum hat.

Die in Kapitel 6 vorgestellte Studie konnte keine Verbesserung der Lernerfolge durch den Einsatz des nicht-muttersprachlichen Englisch messen. Zwar zeigte die Studie, dass sowohl tschechische als auch italienische Teilnehmer die Stimme eines Agenten, der die gleiche nicht-muttersprachliche Varietät spricht konsistent als wärmer und sympathischer einschätzen, aber die Stimme eines Agenten der amerikanisches Englisch spricht als kompetenter empfunden wird. Eine ähnliche Studie in China wird derzeit durchgeführt. Diese Ergebnisse können wichtige Hinweise für das Design und die Implementation ähnlicher Systeme in hybriden Gesellschaften geben um die Akzeptanz derartiger Technologien zu erhöhen und eine reibungslose Interaktion zu ermöglichen.

A.4 Diskussion

Bei der Interpretation der Ergebnisse der hier vorgestellten Studien ist es wichtig (1) die Einschränkungen der Daten und (2) die theoretischen und methodischen Einschränkungen zu bedenken. Die Aufnahme der Daten wurde durch die COVID-19 Pandemie beeinträchtigt und verzögert (siehe Kapitel 1.6.2). Durch die weitreichenden Einschränkungen von sowohl internationalen Reisemöglichkeiten

[1] Umgangssprachlich spricht man häufig von kurzen und langen Vokalen, was jedoch nur die Vokalquantität beschreibt, die Vokalqualität wird duch gespannt-ungespannt beschrieben.

als auch der Möglichkeit Studien mit Teilnehmern vor Ort durchzuführen konnten die Daten erst wesentlich später als geplant und auch nur von Kollegen vor Ort gesammelt werden. Obwohl die Datenaufnahmen so repräsentativ wie möglich geplant waren, stellen sie gleichzeitig eine willkürliche Stichprobe (*convenience sample*) dar, da die Auswahl der Teilnehmer von Kollegen vor Ort getroffen wurde. Das führte dann auch dazu, dass teilweise signifikant mehr Daten weiblicher Teilnehmerinnen vorhanden sind, aus dem einfachen Grund, dass an dem entsprechenden Partnerinstitut mehr Studentinnen eingeschrieben sind. Weiterhin wurden alle Teilnehmer an Universitäten rekrutiert. Das bedeutet, dass die Sprecher alle eher gebildet sind und Englisch ausschließlich in formalen Kontexten benutzen.

Den Studien liegen außerdem eine Reihe theoretischer Annahmen zugrunde, die bei der Interpretation bedacht werden müssen. Die Untersuchung des Einflusses einer Muttersprache (z.B. Mandarin) auf eine Fremdsprache (in diesem Fall Englisch) setzt natürlich voraus, dass die zugrundeliegende Theorie des Fremdspracherwerbs dies zulässt. Die hier präsentierte Analyse des Sprachrhythmus erfordert die Annahme, dass Sprachrhythmus ein gradierbares Phänomen ist und die Einteilung von Sprachen auf einem Kontinuum mit eher silbenzählend und eher akzentzählend an den jeweiligen Enden erlaubt.

Weiterhin ist es schwierig, isolierte Variablen experimentell zu messen und alle möglichen konfundierenden Variablen auszuschließen. Insbesondere die in Kapitel 6 vorgestellte Studie ist von dieser Schwierigkeit betroffen, weshalb folgende Maßnahmen zur Kontrolle konfundierender Variablen ergriffen wurden: (a) der Einfluss von Vorwissen sollte dadurch ausgeschlossen werden, dass ein Text der auf der einen Seite relevant für das Studium der Probanden ausgewählt wurde, jedoch ein neues Thema einführte; (b) die Komplexität der verwendeten Sprache des Materials, welches aus einer wissenschaftlichen Veröffentlichung stammt, wurde mit verschiedenen Textkomplexitätsmetriken analysiert, damit die Sprache fordernd, aber nicht überfordernd ist; (c) der Text wurde recht kurz gehalten um einen negative Einfluss von unterschiedlich langen Aufmerksamkeitsspannen und unterschiedlichen Gedächtniskapazitäten auszuschließen; (d) alle verwendeten synthetischen Stimmen hatten das gleiche Geschlecht (weiblich) um sozial konditionierte Vorurteile konstant zu halten; (e) der Lerninhalt wurde lediglich als Audiodatei abgespielt, d.h. ohne ein Transkript oder andere visuelle Medien, um einen Einfluss des visuellen Kanals aufs Lernen auszuschließen; (f) der eingesetzte Agent hatte kein äußeres Erscheinungsbild, sondern nur eine Stimme, um Vorurteile Lehrern gegenüber zu vermeiden.

Die hier dargestellten Einschränkungen müssen bei der Interpretation der Ergebnisse bedacht werden, schränken die Verwertbarkeit der Ergebnisse jedoch nicht ein. Insbesondere die angesprochenen methodischen Einschränkungen resultieren aus der grundlegenden Natur der sozialwissenschaftlichen Forschung. Deshalb muss in derartigen Studien immer ein Kompromiss zwischen der Kontrolle möglicher konfundierender Variablen und der Natürlichkeit des Experiments gefunden werden.

A.5 Schlussbetrachtung

Im Allgemeinen kann davon ausgegangen werden dass ein nicht-muttersprachlicher Agent die Lernergebnisse Studierender nicht signifikant verbessert, jedoch als wärmer und sympathischer eingeschätzt wird als ein muttersprachlicher Agent. Die detaillierten quantitativen Analysen haben gezeigt, dass das chinesische Englisch von Sprechern aus Guangdong, Guangxi und Yunnan klare nichtmuttersprachliche Charakteristiken, wie einen schwach ausgeprägten Vokalkonstrast und einen eher silbenzählenden Sprachrhythmus besitzt. Weiterhin konnte gezeigt werden, dass insbesondere die Vokalpaare FLEECE-KIT und GOOSE-FOOT nicht unterschieden werden. Die Analyse des Sprachrhythmus zeigte

erstmals, dass tschechisches English eher silbenzählend und italienisches Englisch eher akzentzählend sind. Auf methodischer Ebene haben die Studien in dieser Arbeit einen Beitrag dazu geleistet, automatische Transkriptionen und automatische Analysemethoden als valide Methoden der quantitativen Variationslinguistik zu etablieren. Weiterhin hat insbesondere die vorgestellte linguistisch motivierte Qualitätsmetrik für neurale Sprachsynthesesysteme gezeigt, dass Methoden aus der Linguistik nicht nur genutzt werden können, um die Natürlichkeit der synthetisierten Sprache zu beurteilen, sondern auch Einblicke in Möglichkeiten der weiteren Optimierung des Systems bieten.

Es liegt in der Natur der sozialwissenschaftlichen Forschung, dass die Ergebnisse nur unter Beachtung der auf die Studie zutreffenden Einschränkungen interpretiert werden können. Die in dieser Dissertation vorgestellten Studien haben zum einen theoretische Einschränkungen, indem sie annehmen dass die Muttersprache einen Einfluss im Fremspracherwerb hat und dass Sprachrhythmus ein gradierbares Phänomen ist. Zum anderen schränkt die Art der Datenerhebung die Allgemeingültigkeit der Ergebnisse etwas ein. Da die Daten ausschließlich an Universitäten erhoben wurden, ist die von den Daten abgebildete Sprache nahe an der Standardsprache, da die Verwendung der Standardsprache im Allgemeinen mit Bildung korreliert. Zuletzt gilt es zu beachten, dass es in einer Studie, die Lernerfolge misst eine große Herausforderung ist, alle möglichen konfundierenden Variablen zu kontrollieren. In der vorliegenden Studie wurden daher die Stimuli kurz gehalten und ohne visuelles Material zur Unterstützung dargestellt, um den Einfluss unterschiedlicher Aufmerksamkeitsspannen und Gedächtniskapazitäten, sowie einen Einfluss des visuellen Kanals auf die Ergebnisse auszuschließen. Leider senken diese Entscheidungen die Natürlichkeit der Situation, in der der Lernerfolg gemessen wird. Abschließend ist festzustellen, dass die Ergebnisse der Studien trotz der benannten Einschränkungen wertvolle Ergebnisse geliefert haben und damit die Forschung vorangebracht haben. Aus den Einschränkungen lassen sich Möglichkeiten zu weiterführenden Untersuchungn ableiten.

So wäre es wünschenswert, wenn zukünftige Studien Daten aus anderen Teilen Chinas analysieren, um ein umfassenderes Bild des chinesischen Englisch zu erhalten. Weiterhin wären ähnliche Beschreibungen der Vokalräume des tschechischen Englisch und des italienischen Englisch sinnvoll. Eine Ausweitung der Analysen auf geschriebene Sprache wäre gleichermaßen sinnvoll und wünschenswert und würde es auch erlauben, den Zusammenhang zwischen gesprochener und geschriebener Sprache und einen möglichen Einfluss der Muttersprache durch den Graphem-Phonem Zusammenhang zu untersuchen. Letztlich wäre eine erweiterte, eventuell multimediale Lernstudie ebenfalls sinnvoll, um den Lernerfolg in natürlicheren Szenarien wie Lernvideos oder Podcasts zu messen.

B Eidesstattliche Erklärung zur Eigenständigkeit

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertation selbstständig verfasst und keine anderen als die ausgewiesenen Hilfsmittel benutzt habe. Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach anderen gedruckten oder im Internet verfügbaren Werken entnommen sind, habe ich durch genaue Quellenangaben kenntlich gemacht.

. . . . . .
Ort, Datum

. . . . . .
Unterschrift