Palacký University Olomouc

Doctoral Thesis

Czech accent in English: Linguistics and biometric speech technologies

---

Author:
Mgr. Jakub F. Bortlík

Supervisor:
Mgr. Šárka Šimáčková, PhD.

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Linguistics section of the
Department of English and American Studies

November 8, 2021

Declaration of Authorship

I, Jakub F. Bortlík, declare that this thesis titled, "Czech accent in English: Linguistics and biometric speech technologies" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed:

_____

Date:

_____

Tutto vanità, solo vanità,
Vivete con gioia e semplicità,
State buoni se potete...
Tutto il resto è vanità.

Tutto vanità, solo vanità,
Lodate il Signore con umiltà,
A lui date tutto l'amore,
Nulla più vi mancherà.

Filippo Neri

But it was fit
that we should make merry
and be glad,
for this thy brother was dead
and is come to life again;
he was lost,
and is found.

Luke 15:32

PALACKÝ UNIVERSITY OLOMOUC

Abstract

Faculty of Arts
Department of English and American Studies

Doctor of Philosophy

Czech accent in English:
Linguistics and biometric speech technologies

by Mgr. Jakub F. Bortlík

This dissertation combines the points of view of linguistics and biometric speech technologies to examine the topic of Czech accent in English. It summarizes some common challenges in foreign-accent rating experiments, it presents the creation of an extensive multilingual data set, and explains the data set's use in a complex foreign-accent rating experiment. The results of the accent rating experiment show that the definition of the rating task can be used to direct raters' attention to distinguish more categorically native speakers from non-native speakers by instructing listeners to rate their confidence about their perception of an accent rather than rating accent strength itself. Furthermore, it turned out that non-native speakers who receive accent ratings similar to those of native speakers also group with native speakers when it comes to accent ratings in adverse listening conditions. The results also show that raters' familiarity with native accents of English can be a predictor of confidence ratings but not of accent-strength ratings, and the data confirm the earlier finding that native language match between raters and speakers lowers accentedness scores. The chapter dedicated to biometric speech technologies identifies recording quality as the main predictor of the accuracy of automatic language identification and automatic speaker recognition. It shows that native language and foreign accent ratings do not necessarily correlate with language identification scores but they can be predictors of speaker identification scores when recording quality is factored out.

Acknowledgments

Even though I am the one whose name is on the cover of this book and I use the first person singular frequently throughout the text, the fact that it actually exists is by no means an individual's achievement. Without any claim to completeness (I guess, at the moment, I should put more effort into finishing the thesis itself and later thank the people I forget in person) and in no strictly defined order, here is a list of people who have made this publication possible.

It was Dr. Šárka Šimáčková, my PhD supervisor, who got me into phonetics in the first place and who, since the turn of the twenty-first century, has seen me grow much older and a little wiser under her patient supervision. She helped me to pinpoint my weaknesses and strengths and helped me to stay focused on what mattered by getting rid of the unimportant stuff. Her no-bullsh*t attitude helped me to find the right way beyond the dissertation and my studies at the English Department.

Without the help of Dr. Richard Andrášik from the Department of Mathematical Analysis and Applications of Mathematics, without his assistance with statistics, and his erudition in data analysis, I would have given up just within sight of the finish line. Most of the statistical analyses are his, all statistical mistakes and misconceptions are mine. He helped me decide some questions about data preparation, suggested the statistical methods to use, performed the tests in R, created most of the plots, and provided me with the R code, so that I could prepare some analyses and plots on my own. I am extremely grateful for his invaluable help.

The students of the English Department in Olomouc, and Paul, Mitch, Sarah, and Rachel kindly let their foreign-accented and native voices and metadata become part of my experiment. Lukrécia Sarah Hašková and Silvie Pospíšilová did an amazing job with the data collection (and Martin Schweizer helped a bit too).

I am grateful to Phonexia for letting me use its software for my research and allowing me enough free time to finish the dissertation. Dr. Petr Schwarz of Phonexia and Brno Technical University shared his useful insights on an early draft of the thesis. From time to time, he also expressed his surprise that I still had not finished, which helped me stick to it. Roman Polok supported me the whole time I had a part-time contract and even when I took a month off (twice!) from his team to finish the thesis. His unwavering support, especially during the last weeks, kept me going. Michal Klčo gave me many valuable pieces of advice regarding ASR, LID, evaluations, the SpeechBrain toolkit, and codecs for preparation of the data for the accent rating experiment. And Jirka Nytra prepared for me a special version of the Speaker ID software that could handle some of the limitations of my data set.

All the people who took part in the accent rating experiment deserve my deep gratitude, especially those who did the "telephone" version, which, as one participant put it, "was brutal". Ms. Nicola S. Karásková from the English Department of the Technical University of Liberec, Czech Republic, was especially active in mobilizing her students, and virtually doubled the number of participants in my experiments.

Professor Gijsbert Stoet created the free PsyToolkit platform that I used for the accent rating experiment, and which made it possible to collect data globally even in the time of a pandemic. I am grateful to the multitude of geeks who created all the free software I used when working on the dissertation: Praat, Python, R, GNU/Linux, LaTeX, Vim, and much more. Vel (2017) created the template I used to typeset this thesis.

Dr. Jonáš Podlipský was my mentor at the English Department. He got me started with Praat scripting, which brought back memories of my flirtation with Basic in third grade, and which has since evolved into a passion for free software, and in a way got me my job at Phonexia. He has always been someone to look up to when it came to scientific and academic drive, and someone to look up for a good chat about life, work, and stuff when I was visiting Olomouc.

The head of the Department, Professor Ludmila Veselovská, gave me the brilliant idea to take a month off to finish the dissertation (except I had to take a month off twice, prolonged the studies by a couple of years, moved to a different country and back, got married, found a proper job, and fathered a daughter in the meantime).

Professor Joseph E. Emonds taught me that true learning happens outside of one's comfort zone. I am grateful to him for modelling intellectual audacity with his and professor Faarlund's book *English: The language of the Vikings*. He also kindly offered to let me camp in front of his house after telling him that I could not take a month off of teaching English classes because I had rent to pay. And he wrote me a fabulous letter of recommendation when I was applying for a scholarship with the OeAD so that I could explore a Viennese cul-de-sac in my dissertation journey. (He also explained how to fold Scotch tape so that it can be used as double-sided tape.)

Dr. Michaela Martinková reminded me that "it does not have to be perfect" and took care of the PhD studies agenda at the Dept. of English and American studies.

I should also thank Marcin Wągiel [m a r tɕ i w̃ v ɔ ŋʲ gʲ ɛ l] (almost rhymes with "grotżagiel"). He mentioned me in his book and he would be sad if I did not mention him in mine. He's the best linguist I've ever played basketball with, and the best basketball player among linguists that I know of (maybe after Shaq).

Master of Divinity Graham Kervin proofreaded most of the dyserrtation and made it significantly less painful to read (Kolmogorov-Smirnov, $p < 0.0001$). Any remaining linguistic oddities are my own. Apparently, I have a problem with "which" or "that", I don't know that. But at least now I know which...

Františka taught me how to finish a dissertation and still (try to) be a good dad.

My wife stayed calm and supportive even when I was describing my thesis as a piece of sh*t that I've already sh*tted out once and was eating and sh*tting it out again. Most importantly, she said she would love me even if I did not finish the PhD, and loved me even when I was trying to.

Jesus Christ was showing me the way how not to jump out of the window when this bloody thing was driving me crazy.

Contents

List of Figures

List of Tables

List of Abbreviations

ACC	**A**ccuracy
ASR	**A**utomatic **S**peaker **R**ecognition
BSAPI	**B**rno **S**peech **A**pplication **P**rogramming **I**nterface
Cllr	**L**og-**L**ikelihood-**R**atio **C**ost
DET	**D**etection **E**rror **T**rade-off
ECAPA	**E**mphasized **C**hannel **A**ttention, **P**ropagation and **A**ggregation
ECDF	**E**mpirical **C**umulative **D**istribution **F**unction
EER	**E**qual **E**rror **R**ate
FAR	**F**oreign **A**ccent **R**ating
FA	**F**oreign **A**ccent
FNR	**F**alse **N**egative **R**ate
FPR	**F**alse **P**ositive **R**ate
GA	**G**eneral **A**merican
ID	**Id**entification
KS	**K**olmogorov-**S**mirnov (statistical test)
L2	**2**nd Language
LID	**L**anguage **Id**entification
PVI-V	**P**airwise **V**ariability **I**ndex of **V**owels
RP	**R**eceived **P**ronunciation
RR	**R**isk **R**atio
SID	**S**peaker **Id**entification
TDNN	**T**ime **D**elay **N**eural **N**etwork
VOT	**V**oice **O**nset **T**ime

Pro Lucinku, dugongy a ostatní pozemšťany.

1 Introduction

In this thesis, I present a unique combination of foreign accent research and speech technology evaluation. First, I report a number of experiments I conducted in order to examine several topics related to foreign accents: I had a look at some questions of foreign accent research methodology regarding different ways of formulating foreign-accent rating tasks. Then, I investigated foreign accent ratings in adverse listening conditions and correlations with rater's native language and familiarity with English accents. Finally, I tried to find out how foreign-accented data affected the accuracy of selected biometric speech technologies: language identification and speaker recognition.

Speakers of a language as their *second language* (i.e., other than their *native* or *first* language) can often be distinguished from native speakers by differences in pronunciation. These are usually referred to collectively as a *foreign accent*, especially if the different pronunciation features are caused by interference with the native language. Foreign accents in non-native speakers have been studied extensively from a linguistic point of view (e.g., see (Edwards and Zampini 2008) on phonology and second language acquisition, and (Moyer 2013) for an applied-linguistics account). Far less frequent has been the application of the findings of foreign-accent research in speech technologies. Generally speaking, speech technologies are computer programs based on neural networks and other machine learning techniques that process natural speech recordings in order to make information about the speaker and the content of the recordings available to technology users. The influence of foreign accents on speech technologies poses many interesting questions. Answers to these questions could help researchers and developers improve their software, as well as help its users better understand its possibilities and limitations.

In the globalized world, foreign-accented speech is more frequent than ever before. According to Omniglot (2021), English currently ranks third with respect to the number of native speakers, totalling around 350 million, but it is the most widely used second or foreign language, with estimates reaching 850 million non-native speakers worldwide. In the United States alone, English may be a second language for more than 55 million people who have reported speaking a different primary language at home.[1] In such a world, speech technologies are bound to be applied to non-native speech rather frequently, and technology users as well as technology providers

1. To bring this number into perspective, English is the primary language at home for slightly over 231 million people living in the U.S. (Worldatlas 2021).

may benefit from the knowledge of how well speech technologies can perform when encountering foreign-accented data.

I was inspired to do this research by my previous work on glottalization in Czech and English and by my current job at Phonexia, a company based in Brno, Czech Republic, which develops a range of speech technologies that are used commercially in cases such as contact centers, chatbots, and voice commands for human-machine interaction. They are also deployed in various law-protection and security contexts, such as forensic speaker recognition, police investigations, fraud detection by financial institutions, and in crime-fighting and counter-terrorism operations performed by government and intelligence agencies.

The range of technologies developed by Phonexia includes identification technologies (Speaker Identification, Language Identification, and Gender Identification), and technologies that process the content of what is being said (e.g., Speech Transcription, Keyword Spotting, and Time Analysis, which estimates speech rate and measures turn-taking patterns in two-channel recordings).

The technologies are trained primarily on monolingual[2] native-speaker data from telephone calls. However, in real-world applications, the software is also applied to non-native, foreign-accented speech, as well as data coming from different sources, such as a variety of voice recorders, mobile devices, and other microphones.

The type and degree of foreign accent could influence the accuracy and/or type of error in speech technologies trained primarily on native-speaker data. Speakers could be incorrectly matched by speaker identification technologies, languages incorrectly classified by language identification, or words recognized incorrectly by automatic speech recognition due to differences of second language phonetics or phonology from sound patterns in native speech.

In order to conduct the experiments, I collected a corpus of studio recordings of English spoken by Czech and Slovak native speakers, and with native English speakers and a Ukrainian/Russian native speaker as controls. I prepared a parallel data set which simulated the quality of telephone recordings. Then I collected foreign accent ratings on both subsets from speakers of various languages from around the globe using the online platform PsyToolkit (Stoet 2010, 2017). Then I processed the data set with SpeechBrain and Phonexia speech technologies and compared the results with their foreign-accent ratings. Finally, I "tortured the data until it confessed", as the saying goes. Let's see how it was all done and what came out of it.

2. By monolingual I mean that training recordings contain only one language, and that individual speakers usually provide data in one language.

2 Overview of foreign accent research

Existing studies on foreign accents have various kinds of motivation. Some people use the research results to enhance English as a Foreign Language courses, e.g., in pronunciation training and accent reduction as a part of a curriculum, or to improve English proficiency testing. Others try to expose foreign accents as a cause of discrimination and bias against non-native speakers (e.g., Sato 1991; Stocker 2017). Still others strive to tease apart factors of age of learning onset, length of learning, intensity of instruction, native language, and language learning aptitude in the development and retention of a foreign accent. Sometimes this happens in the pursuit of a definition or refutation of a *critical period* hypothesis for language acquisition, sometimes out of pure scientific curiosity, and sometimes, unfortunately, in order for authors to publish lest they perish. As this thesis is presented "in partial fulfillment of the requirements for the degree of Doctor of Philosophy", as the phrase goes, a certain measure of "lest the author perish" is unavoidable . But hopefully the reader will also find glimpses of real curiosity and of a true desire to better understand how things work with foreign accents, noisy data, and speech technologies.

In current research, the distinction is generally made between the foreign accent, intelligibility, and comprehensibility. While in everyday language, intelligibility and comprehensibility are synonymous (meaning "the fact of being able to be (easily) understood", Oxford Learner's Dictionary), in foreign accent studies, the terms are used differently to keep separate the objective and subjective points of view. *Intelligibility* can be defined as "the extent to which a speaker's message is actually understood by a listener" (Munro and Derwing 1995, 289), while *comprehensibility* ratings are based on the opinions of listeners who are asked to make judgments of *perceived* comprehensibility (287). In contrast to opinion-based comprehensibility, intelligibility can actually be measured by having listeners transcribe speech samples and by comparing those transcriptions with the reference texts.

Numerous studies have found that the perceived degree of foreign accent, intelligibility and comprehensibility does not only depend on the speaker and the characteristics of the rated speech samples themselves. On the contrary, there seems to be a complex relationship between factors relating to the *speaker*, the speaking and listening/rating *task*, the rated *material*, and the *listener*.

There are many aspects of rating experiment design in which existing studies differ. Jesney (2004) provides a rather exhaustive overview of studies published until

2004, and summarizes the various aspects. In the following overview I present just a few that seem relevant to my research. I try to deal mainly with those aspects of foreign accent rating (FAR) that have some bearing on speech technologies and the data to which speech technologies typically apply.

In this thesis, I will use the terms *native listener* and *native rater* to denote "a person whose task it is to listen to and rate the foreign accent of a sample produced in their native language". Similarly, the term *native talker* will be used instead of *native speaker* in the sense of "a person who is the native speaker of the language that they are speaking in a sample". I understand that "native talker" is not the most natural expression, as it sounds a little like "a born talker" (i.e., someone who can talk well), but I chose this designation in order to highlight the *action* of a native speaker *speaking*, as well as to avoid other more explicit yet awkward phrasings.

2.1 Native language (mis)match

In (Wester and Mayo 2014), native listeners have been found to give lower accentedness ratings to native-talker samples in their own language than the ratings given by non-native listeners (in other words, even some native talkers can sound more foreign-accented to non-native listeners). At the same time, the study found that in rating samples produced by non-native talkers, native listeners gave higher accentedness ratings than did non-native listeners, i.e., native raters were harsher with the non-native talkers. These conclusions corroborated earlier findings by Flege (1988) who focused on the effect of familiarity with the rated second language (L2) in non-native listeners. Flege found that non-native listeners were able to gauge foreign accents qualitatively in a manner similar to native listeners (i.e., they gave higher accentedness scores to non-native than to native talkers), however, the non-native listeners' ratings were less extreme, more average, for both talker groups. Flege's study also suggested that more experience with the target L2 in non-native listeners contributed to more native-like accentedness ratings. That means, L2 listeners with more experience with the target language gave lower accentedness ratings to native talkers and higher scores[3] to non-native talkers than did less experienced L2 listeners (74).

Wester and Mayo (2014), replicated the finding that non-native listeners can give *qualitatively* similar accent ratings to those of native listeners, but can differ in the *range* of ratings they assign. The raters in the study were less harsh with non-native talkers and stricter with native talkers.

In contrast to (Flege 1988), Wester and Mayo (2014) also investigated how the listeners' native language interacted with the talkers' native language when rating both native and non-native English utterances. They found that English native speakers were the listener group that gave the lowest accentedness ratings to English native talkers and, at the same time, gave higher accentedness ratings than any other listener

3. On a terminological side note, "higher score" could be interpreted as "better", i.e., more "native-like" pronunciation, however, I use the term "high score" as equivalent to "high accentedness rating", i.e., "strong accent".

group to all three non-native talker groups (Finnish, German and Mandarin speakers of English). However, the findings for non-native listeners did not follow this pattern. Non-native listeners did not automatically judge less harshly the foreign accent of non-native talkers who shared the native language with them. E.g., German listeners gave higher accentedness ratings to German speakers of English than did Finnish and Mandarin listeners. Including both native and non-native raters (with and without the matching native language) seems to provide a better overall picture of the accents included in the rating data set.

Bent and Bradlow (2003) demonstrated what they called the "interlanguage speech *intelligibility* benefit" (emphasis mine). They found that for non-native listeners, highly proficient non-native talkers (with either the same or different native language as that of the listeners) were at least as intelligible as native talkers. Based on (Wester and Mayo 2014) mentioned above, there does not seem to be a similar "matched-interlanguage *accent-rating* benefit". In light of the finding in (Schmid and Hopp 2014) that the reference points of the rating scale and the way the rating task is formulated both influence FA ratings, it can be hypothesized that the various second-language rater groups in (Wester and Mayo 2014) might have had different ideas of what constitutes a foreign accent, which was the reference point in the study. Their results could have been different had the rating task been formulated in terms of "nativeness" rather than "foreignness".

2.2 Accent familiarity

While quite a few studies have analyzed the effects of language familiarity on listener ratings, the effects of *accent* familiarity have received comparatively little attention. Moreover, studies have mostly investigated the effect on comprehensibility or intelligibility, not on accentedness ratings. Schmid and Hopp (2014) found that nativeness judgments were lower for raters with a lower degree of familiarity with the target language and what they call the "contact language" (which is probably to be understood as language with a foreign accent). In contrast, Huang (2013, 783) "did not find any significant effects of accent familiarity ... on raters' evaluations of non-native speech". This discrepancy may have to do with the difference in focus on either nativeness or accentedness (see Section 2.3), and with the exact meaning of accent familiarity: familiarity with foreign-accented speech in general, or familiarity with foreign accent based on a specific native language.

Intuitively, I would expect that good familiarity with native accents would cause raters to give low accentedness ratings to native talkers and higher accentedness ratings to non-native talkers. On the other hand, if raters self-report high familiarity with a particular foreign accent, it might be that they are themselves speakers of that accent and, in such a case, their familiarity may come from what they are used to hearing (imagine a typical Czech high school or university student of English who

is immersed in the Czech-accented English of his or her fellow students and teachers). They may be unaware of how characteristic this accent is[4] and therefore might judge talkers with this accent as less foreign-accented than speakers of another foreign accent, even if this other accent was weaker in the eyes, or rather in the ears, of native listeners. In contrast, a Czech speaker of English who has only had limited experience with Czech-accented English, perhaps because he or she learned English in an English-speaking country, might immediately spot the unique characteristics of Czech-accented English. Similarly, a native speaker of English who is well familiar with Czech accents in English might be better equipped than the naive non-native listener to detect the accent even in the speech of very proficient Czech speakers of English. It seems that accent familiarity should be considered together with familiarity of target language and native language of the rater.

2.3 Rating task

2.3.1 Frame of reference

Existing studies show that foreign accent ratings are influenced by the formulation of the rating task. First of all, the task formulation can determine the *frame of reference*. Some experiments ask listeners to rate how *native* a speaker sounds, other experiments make raters refer to *foreign accent*. In Schmid and Hopp's 2014 view, foreign accent does not have a defined "standard", in contrast to a defined native standard. Schmid and Hopp (2014) suggest that people have "a relatively homogeneous implicit standard of nativeness against which they can make proportional judgements of non-native accentedness. However, they differ in their understanding of foreign accentedness." Their suggestion is based on German data, and it seems questionable whether a similar claim can be made about English and other languages that are much more pluricentric than German[5], and with more local standards, e.g., British, American, Australian, Indian, and South African, to name just a few. So for languages like English, the notions of foreign accent and native speaker accent may be equally vague.

2.3.2 Focus

The task formulation can also determine a rater's *focus* on what is actually being rated. In some experiments, raters are asked to rate the *strength of foreign accent* in audio samples; in others, raters focus on their *confidence about the presence of a foreign accent*. Some studies ask both questions about each sample: participants may first

4. I recall the conversation of a group of high school students on a train, one of whom claimed that English pronunciation is very easy, in contrast to some other foreign language. As likely as not, the student was only referring to some basic understanding of how English orthography relates to its pronunciations but was not aware of the niceties of English phonetics and phonology, such as aspiration of voiceless stops in front of stressed vowels, missing aspiration of voiceless stops after /s/, "dark" and "clear" /l/, diphthongization of long vowels in some varieties of British English, etc.

5. German itself is not a particularly uniform language, with local standards existing in Austria, Germany, and Switzerland, among other regions.

have to rate the strength of foreign accent in each sample or how native a speaker sounds, and then express their confidence about the rating (e.g.,). Raters' confidence about the presence of foreign accent (or about their rating) and the strength of the accent itself seem to be quite different things, and they could influence raters to differentiate between native speakers and non-native speakers more clearly—maybe even those non-native speakers who have only a weak accent.

Some forms of foreign accent can certainly be detected with high confidence by human listeners even if the accent is weak. For example, some Polish speakers of Czech can have stress patterns or use vowel lengths that deviate from the native norm; some Slovak speakers of Czech can be spotted by their retroflex pronunciation of post-alveolar fricatives and affricates, cf. (Hanulíková and Hamann 2010) and (Šimáčková, Podlipský, and Chládková 2012). This would most likely lead to different ratings for the same weakly accented sample if the task were to rate either accent strength or the rater's confidence about the presence of the accent. It is unclear whether such kinds of weak foreign accent would also have a significant effect on speech technologies, or whether only forms of foreign accent perceived as strong by humans (possibly also with a high degree of confidence) would have any significant influence on speech processing software. Strongly accented samples, on the other hand, might be likely to elicit both high accent strength and high rater confidence ratings.

2.3.3 Rating scale

Another important topic is the way in which raters express their opinion about the rated material. The overview in (Jesney 2004) shows that this has typically been done by having raters select points on a Likert scale. Likert scales generally have a number of equally spaced levels with contrasting labels at each endpoint. A minority of studies used sliding scales with no markers along the scale or a marginal method called "direct magnitude estimation", which has been found to produce data equivalent to Likert-scale ratings and does not need to be discussed here. The problem with Likert scales is that they produce ordinal data—the scale points are ranked but, as Jamieson (2004) puts it, "the intervals between values cannot be presumed equal", and as such, Likert scale measurements should not be analyzed using arithmetic means and standard deviations. Nevertheless, this practice has been commonplace in foreign accent research from early on up to more recent studies: Olson and Samuels (1973), Anderson-Hsieh and Koehler (1988), Flege, Frieda, and Nozawa (1997), Munro, Derwing, and Morton (2006), Schmid and Hopp (2014), and many others all use arithmetic means (sometimes complemented by standard deviations) and different flavors of ANOVA for analysis of foreign accent ratings. The same problem has been pointed out in the field of grammaticality ratings elicited on Likert scales because of the difficulty of fitting linear regression models to ordinal data, and also because of the limitations of freely available statistics software to perform ordinal regression with several random-effect factors (Baayen 2008).

2.4 Raters' awareness of data set composition

Studies differ in how much raters know about the material they are going to hear. Some experimenters inform participants beforehand about the native language of the non-native English speakers: Volín and Skarnitzl (2010), for example, explicitly instructed raters "to mark the degree of *Czech* accent in each utterance" (emphasis mine). This is possible when it can be expected that raters will be familiar with this kind of accent and, of course, when the speakers actually do have this accent. It would be problematic to ask listeners to rate a particular non-native accent if the data set contained native talkers or native speakers of another language.

Some studies explicitly inform participants about the presence of native speakers in the data set and remind them of the natural regional variation of native speech. Wester and Mayo (2014), for example, instructed subjects "not to consider any ... native English accents as foreign". This can influence raters' frame of reference (cf. section 2.3.1) and boost their existing familiarity with accent variation (cf. section 2.2).

2.5 Speaking style

Speaking style is known to influence many aspects of pronunciation both on the segmental and suprasegmental level. In phonetic research, several speaking styles are commonly distinguished. The "disorderliness" of the field in the past can be seen in Llisterri's 1992 somewhat tedious overview of commonly used terminology and proposal for further (probably similarly tedious) papers which should establish a "standardization of labels and definitions" in the speaking-style field of research. In all the confusion presented in (Llisterri 1992), the distinction between *read* and *spontaneous* speech still seems to make sense and is used in contemporary research. For example, Spilková and Dommelen (2010) found that some word forms were reduced more in spontaneous speech than in read speech. Kuschmann and Lowit (2015) believed they found that unscripted speaking styles (i.e., semi-spontaneous "picture description and a monologue") showed wider variation of intonation, whereas scripted speaking styles (i.e., read "short sentences and a text passage") more clearly reflected differences in the functional use of intonation contours.[6]

When it comes to the distinction between read and spontaneous speaking styles, some linguists might argue that only spontaneous speech is an example of natural language usage, and that reading, since it is rooted in writing, is irrelevant to the study of natural language. Still, there are good reasons why phonetic studies often rely on read speech.

6. Indeed, in informal observations of Czech I've noticed examples of people trying to make "functional use of intonation contours": It is not uncommon that people (even individuals with some training in public speaking) use rising intonation when they read a sentence ending in a question mark, even if it is a *wh*-question that normally requires falling intonation to sound natural or like a genuine inquiry, and not just a rephrasing of someone else's question.

One reason is that with reading it is much easier to control segmental features, length, and fluency in the studied material, even if it is "unnatural" in comparison with spontaneous speech. This seems to me a little like studying *Arabidopsis thaliana* and *Drosophila melanogaster* as model organisms in biology and genetics. These organisms can be easily cultured and their research can provide insights into other more complex organisms and phenomena (Ng et al. 2018). To maintain control over segmental features, fluency, etc. in a large dataset of spontaneous speech samples would be extremely time-consuming, if at all possible.

Another reason why the read speaking style deserves attention in research of accent rating and speech technologies is that "speech type" is a factor in forensic voice discrimination (Smith et al. 2018), and, as such, its analysis has application and serious implications in real life. Read speech can be used by forensic experts in examinations of case material and can be used in court either to put people in jail or to set them free. This is certainly a good reason why the influence of speaking style on the performance of both human and automatic speaker recognition (ASR) is not just a fun intellectual exercise but should be well understood.

While spontaneous speech is considered by some the "right kind of natural language" and deserves our ultimate attention, it presents serious limitations for data-driven phonetic research, including accent rating experiments. It may be far more difficult than with read speech to find spontaneous speech samples comparable with respect to length, fluency, grammaticality, phonetic and lexical content. This can be a problem if one wants to compare accent ratings of such data or to analyze (as an experiment I recently took part in does) how "sexy", "beautiful", and "structured" a language sounds, or what its "social status" is.[7] On the other hand, for the same reasons it can be easier to collect corpora of spontaneous speech that have more varied and "random" content, which is the type of data that can be used for training new models in some speech technologies, like text-independent speaker recognition.

2.6 Stimulus length

Existing studies vary greatly in the size of samples that are rated for foreign accent. They use either various units of speech (syllables, words, phrases, sentences), paragraphs of text, or timed clips (from 30 milliseconds to two minutes). The interested reader can find a detailed overview of studies and the stimuli length they used in (Jesney 2004).

Flege (1984) found that phonetically trained, as well as naïve American listeners, could detect a French accent in English in extremely short samples, even in just the release phase of the obstruent /t/. On the other hand, Volín and Skarnitzl (2010)

7. In the meantime, the experiment at https://phonaesthetics.de may have been taken down and there may be a published study based on it that may reveal that the purpose was not to study the "sexiness, beauty, structuredness and status", but something completely different. Nevertheless, the point is that the experiment used equivalent excerpts from *The North Wind and the Sun* in various languages.

obtained "almost identical mean ratings" for short paragraphs of 70–90 words and for short utterances of 12–14 words, and considered the paragraphs to be "excessive" and "quite costly in terms of research time and perhaps unnecessary" (Volín and Skarnitzl 2010, 1012).

In light of the findings in (Flege 1984) mentioned above, it seems that the degree of accent can vary within a single word. It seems that if a whole phrase is rated for foreign accent, raters need to average the degree of accent in their head "on the fly". It might be useful to look at the variation that can be found within individual utterances and see what causes the perception of foreign accent in individual sounds, but since the perception of foreign accent is based on both segmental *and* suprasegmental features (see Pellegrino 2012, for example), it seems that one cannot simply dissect a phrase, measure the degree of accent for individual parts, and add up the results. Instead, it seems necessary to select stimuli that enable evaluation of both segmental and suprasegmental features.

Another reason why full paragraphs or long clips are problematic as rating stimuli is that long samples exhaust raters if the corpus is large. In the case that there are just a handful of speakers producing one or two phrases each, this is not a serious problem. In a research project that takes seriously the need for enough data for statistical evaluation, though, individual samples should not be too long at the expense of having only a very few talkers.

Short phrases thus seem to best enable raters to evaluate suprasegmental features such as intonation and speaking rate. At the same time, short phrases seem appropriate for producing more reliable ratings because they give listeners the ability to rate a compact unit (and therefore do not have to average accent features from too long a stretch of speech).

2.7 Foreign accent in adverse listening conditions

Speech samples captured in soundproof booths enable researchers to study very fine acoustic details of pronunciation but they create artificial "in vitro" data that can lead to findings that do not necessarily transfer to natural speech. In real life, people seldom speak and listen to speech in complete silence. Similarly, speech technologies are frequently applied to low-quality audio data coming from people who dictate to their smartphones in the streets, select radio stations or input navigation in their cars by voice, or allow that same voice to be authenticated by the bank over the telephone.

There is a type of research that focuses on how noisy stimuli affect people's perception of speech. This is usually referred to as *adverse listening conditions*.[8] Existing studies show that acoustic conditions can affect how the degree and type of foreign accent is perceived by human listeners (e.g., Lecumberri, Cooke, and Cutler 2010).

8. The flip side of the coin is the fact that acoustic conditions also affect the way people produce speech. The so-called "Lombard effect" describes the fact that people put more effort into articulation if they are surrounded by noise (e.g., Marxer et al. 2018). This is a topic that I will not go into in this thesis but would be a natural complement to research on adverse listening conditions.

Audio quality also affects the performance of automatic speaker recognition systems (e.g., Künzel and Alexander 2014; Morrison and Enzinger 2016). What is unclear is how audio quality interacts with foreign accents in their influence on the performance of speech technologies.

Audio quality could affect some aspects of foreign accent more than others. For example, the realization of the English dental fricative /θ/ as an alveolar [s] or a labio-dental [f] (frequently encountered in foreign-accented speech) will most likely not be distinguishable from an actual dental sound [θ] telephony recordings. In this kind of data, usually "the usable voice frequency band ranges from approximately 300 Hz to 3400 Hz" (Institute for Telecommunication Sciences 1996), and while [s] has its spectral peaks located in a frequency range of 3.8–8.5 kHz, [f] and [θ] were found to have a diffuse spread of energy from about 1.8–8.8 kHz (Behrens and Blumstein 1988). This would make them very hard to distinguish in telephone recordings even in native speech. On the other end of the spectrum, filtering of the lower frequencies affects pitch perception and thus could mask deviations from nativelike intonation patterns. However, the perception of pitch is not based solely on the fundamental frequency. It also relies on ratios between harmonics (cf. Oxenham 2012), so pitch can theoretically be recovered even from filtered telephone speech, although the perception of intonation in full-range and filtered bandwidth is likely to be easier or more nuanced.

Volín and Skarnitzl (2010) found that accent rating differences were gradually equalized with worsening listening conditions, e.g., samples with harsher listening conditions (more coffee-shop noise, more frequency filtering) received more average accent ratings for both highly proficient Czech speakers of English and those with a heavy Czech accent in English. However, even in degraded listening conditions, some prosodic features were robust with respect to predicting FA ratings, or even more robust than in clean listening conditions. Specifically, the study found that the pairwise variability index of vowels (PVI-V[9]), F0 standard deviation, and the declination trend of F0 showed the strongest correlations with FA ratings under less favourable listening conditions. However, they also observed what might be an interplay of segmental and suprasegmental features. Even though rhythmic characteristics (PVI-V) were, in general, found to be predictors of FA ratings, in the harshest listening conditions of one experiment (frequency filtering), some distinction between the three otherwise different accent groups was lost—even though the rhythmic characteristics were retained in the filtered speech. Based on this finding, the authors hypothesized that the rhythmic characteristics apparently interact with some segmental features that are not preserved in filtered speech.

In contrast, the negative correlation between articulation rate (measured as the number of syllables per second and as the number of phonemes per second) was found to become weaker with worsening listening conditions. Faster speakers were found to receive better accent scores (i.e., sounded more native-like), however, this correlation

9. The PVI-V index is a metric of rhythm which shows the level of variability in successive measurements of vowel durations, and the variability in measurements of the duration of intervocalic intervals (Grabe and Low 2002).

was weaker in filtered speech. The authors linked this to the fact that in filtered speech all speakers were rated as less accented than in clean speech.

There is also the question of how to practically collect a data set of foreign accent ratings in adverse listening conditions large enough to allow for statistical analysis yet of good enough quality to make any sense at all. Some researchers, utilizing large groups of students in auditoriums, surrender much of their control over the progress of the experiment, and their listening conditions can vary widely according to the specifics of the auditorium. In contrast, other researchers play samples through headphones to individual raters in soundproof booths, eliciting more qualitatively reliable data but giving up on population sample size. The conflict between quality and quantity does not always lead to a clear solution and compromises are almost inevitable.

3 A brief introduction to biometric speech technologies

Biometric speech technologies analyze speech recordings to get information about speakers' identity, sex, age, and the language used in the audio. This information can be used to improve security and the user experience of services provided by call centers, banks, smart homes, voice control in cars, etc., and biometric speech technologies also find important application in the security sector. They are the subject of extensive research in the academic sector, and they are also being developed by commercial companies. I have the opportunity to work in one such company, Phonexia, to test its software with foreign-accented data, and I compare the results with the speech technologies available in the open-source toolkit SpeechBrain.

3.1 Speaker recognition

Recognizing speakers by their voice has been called "a commonplace of human experience"[10] (Bulejcik *v* R 1996). While the task itself may indeed be an everyday activity, research suggests that if you are not familiar with a voice or if you are not a specially trained expert, recognizing a speaker *correctly* by their voice is not such an easy task after all. Experiments with voice lineups have reported correct identification rates as low as 42 % in some cases and false positive identifications of 51 % in others (Kerstholt et al. 2004). As Edmond, Martire, and Roque (2011, 410) point out, "*false alarms* are commonplace human experiences" (emphasis mine), just like the *task* of recognizing speakers by their voice.

In various situations, however, listeners may be required to perform speaker recognition tasks that can have serious consequences, without having the expertise of a forensic specialist. In some legal traditions, earwitnesses or jury members are asked to recognize speakers in legal proceedings. As another example, police officers without special training prepare material from a case investigation for expert forensic analysis. In recent years, there has been a serious discussion about the logical correctness of procedures of data preparation for automatic Speaker ID systems, where lay listeners, such as police officers without special training, may play a crucial role (Morrison,

10. To put this quote into context: "Recognition of a speaker by the sound of the speaker's voice is a commonplace of human experience. To recognise the voice of a particular speaker some familiarity with that speaker's voice is ordinarily needed (3). A person who is not familiar with the voice of a putative speaker may be able nevertheless to recognise the speaker's voice by comparison with an established example of that voice if the speaker's voice exhibits sufficiently distinct features to permit an ordinary person to identify the speaker or if the person possesses an appropriate expertise" (Bulejcik *v* R 1996).

Ochoa, and Thiruvaran 2012). Furthermore, situations where lay listeners or forensic experts are required to recognize speakers by their voice are frequently complicated by language mismatches, foreign accents, and adverse listening conditions.

3.1.1 Channel and language mismatch

ASR systems are trained on hundreds or thousands of speakers so that they can learn to cope with within- and between-speaker variability. An ideal ASR system should capture the features unique to individual speakers and ignore the effects of language and acoustic characteristics of the audio data.

In ASR terminology, *channel* usually refers to the way in which an audio file was recorded and transmitted, which can significantly affect the acoustic quality of audio files. Channel mismatches, a frequent occurrence in forensic and other equivalent use cases, can therefore have a negative effect on ASR performance (see Morrison, Ochoa, and Thiruvaran 2012). Comparing studio recordings with telephone calls would be a good example of such a mismatch, or *cross-channel* comparison. The opposite kind of comparison is usually called *matched-channel*. In this respect, ASR systems behave similarly to human listeners who are also influenced in their perception of speech by acoustic characteristics of the source (see section 2.7).

Language mismatch, or *cross-language* comparison refers to the situation when two recordings in different languages are compared, as opposed to the so-called *matched-language* comparisons. In personal communication, several forensic experts have told me that they in fact do not analyze samples in languages in which they are not an expert. A German native speaker, for example, would not take on a case that requires him or her to analyze samples in both German and Polish if they are not fluent in Polish because, in the traditional *acoustic-auditory* framework of forensic speaker recognition, an important role is played by sociolectal, dialectal, etc. variations that can only be successfully analyzed with a good knowledge of the given language. Extensive research on the topic of speaker recognition across languages has identified the so-called *language-familiarity effect*, in which "listeners identify voices more accurately in their native language than an unknown, foreign language" (Perrachione 2019). An efficient L2 speaker might even develop a different voice timbre in the second language that would make them even more difficult to recognize as the same person. Forensic experts thus rely on a knowledge of phonological, lexical, and syntactic patterns of the language that are most probably ignored by *automatic* speaker recognition systems.

An ideal ASR system should be language independent—it should correctly recognize a speaker of any language and it should recognize him or her in recordings in different languages. Therefore, the system should extract only the characteristics of a speaker's voice that do not depend on the language. However, L2 speech usually differs from L1 speech. There seem to be some characteristics of L2 speech that apply generally to most speakers regardless of the language combination: L2 speech tends to be overall slower than L1 speech (see references in Aoyama and Guion 2007, 283),

and the L2 speech of less proficient speakers might also contain more and longer pauses (Trofimovich and Baker 2006).

Depending on what kind of features a particular ASR system extracts, it might mistakenly attribute some characteristics to the speaker as a biometric measure, even though they are in fact the result of speaking a foreign language. In a system that does not reliably separate language dependent factors or includes in the speaker model some information about the phonology of the language, a stronger foreign accent, in terms of retaining segmental features from one's L1, could lead to a better matching of recordings of one speaker in L1 and L2. On the other hand, if the foreign accent involved new phonetic forms that deviated from both L1 and L2, the foreign accent would probably have no positive effect on ASR results. Similarly, if the L2 production involved hesitations, filler sounds, longer pauses, and/or a slower articulation rate, these might be confused for speaker characteristics and the foreign accent could then be correlated with more incorrect matching by the ASR system.

In summary, automatic speaker recognition systems have the *potential* of being language independent. They also have the advantage of being relatively easily testable, in contrast to human judgments, which cannot be easily repeated and are much more time-consuming, and are thus virtually unfeasible to the extent made possible by speaker recognition technologies. This brings me to the topic of speaker recognition tasks and evaluation metrics.

3.1.2 ASR tasks and evaluation metrics

Doddington et al. (2000) provide a comprehensive overview of automatic speaker recognition evaluations: here follow just a few points relevant to our topic. In speaker *identification* we compare a recording of an unknown speaker with many recordings of known speakers in a database; a typical example would be a blacklist of known fraudsters in a bank. In the speaker *verification* use case, two voices are compared to see if they belong to the same person, which could be a case of authentication by voice, or forensic speaker recognition.

A speaker verification system can be evaluated by two measures: False positive rate and false negative rate. A false positive, also referred to as "false acceptance" or "false alarm", is a type of error of an ASR system, in which two voice samples are marked as belonging to the same person when in fact they belong to two different people. Conversely, a false negative (also referred to as a "false rejection" or a "miss") is a type of error in which two voice samples are incorrectly marked as belonging to two different speakers. The decision is typically made by comparing a similarity score to a threshold or decision point. In the context of ASR, a comparison is usually referred to as a *trial*.

Both types of errors need to be considered together when establishing the accuracy of a speaker recognition procedure or automatic recognition system. In different use cases, one or the other type of error may be of greater importance. For example, in the context of forensic ASR, either type of error could lead to the incorrect conviction

of a defendant depending on whether a voice recording should prove the guilt or innocence of that defendant. According to the principle of *in dubio pro reo*, any doubts about the guilt of the defendant should be resolved in his or her favor. So if the defendant is supposed to be speaking, for instance, in an intercepted telephone call involving blackmail, then a false positive decision is the more serious mistake. When, in contrast, the defendant is supposed to be speaking in a recording that provides him or her with an alibi, such as a telephone call from a particular landline device, then it is more important to avoid a false negative decision.

The performance of an ASR system can be represented by the *Equal Error Rate* (EER), which combines false positive rates and false negative rates. The Equal Error Rate is measured by adjusting the threshold in an evaluation so that we find the decision point at which false positives and false negatives have the same frequency[11], the so-called *EER threshold*. The lower the EER values, the better the performance. The EER threshold itself is a valuable piece of information that expresses how well the system is calibrated with respect to the evaluation data; a well calibrated system can be expected to have an EER threshold close to 0.

3.1.3 Phonexia `SID4-XL4`

Phonexia has developed several versions of its Speaker Identification technology. The version currently marketed as the most accurate is model `SID4-XL4`. It is based on the x-vector architecture described in (Snyder et al. 2018), and a detailed description of a beta version of `SID4` can be found in (Jessen et al. 2019). The model was trained primarily on telephone data but offers the user several possibilities to increase performance by providing additional calibration and normalization data.

The speaker recognition works in several steps. First, the application uses a deep neural network to analyze the spectral features of an audio file and create a vector that should represent the unique characteristics of the speaker's voice. The vector is saved in a file called "voiceprint". In the next step, the voiceprints extracted from different recordings are compared and the system returns a score in the form of a log-likelihood ratio (LLR), which can be expressed by Equation 3.1.

$$LLR = \log \frac{p(Evidence|Hypothesis\ 1)}{p(Evidence|Hypothesis\ 0)} \tag{3.1}$$

LLR corresponds to the natural logarithm of the ratio between two probabilities: the probability (p) of obtaining a given similarity of two recordings, or *Evidence*, if both recordings come from the same person (*Hypothesis 1*), and the probability of obtaining the same *Evidence* if they come from different people (*Hypothesis 0*). Taking the logarithm of the ratio makes sure that, in a well-calibrated system, the resulting score is centered around zero and can be roughly interpreted as in Table 3.1.

11. We look at the frequency of false positives in the total number of different-speaker trials, and the frequency of false negatives in the same-speaker trials. The frequency of both kinds of errors taken together in the total number of trials can be used as a simple performance measure in the case of speaker identification (Doddington et al. 2000).

LLR	Interpretation
$-\infty$	the system is sure the speakers are different
0	the system is not sure (probabilities are the same)
$+\infty$	the system is sure the speakers are the same

TABLE 3.1: Interpretation of log-likelihood ratio values. In practice, ASR systems use calibration methods so that the LLR values can typically be found in the order of units or tens.

By way of example, if we make a recording of Peter and measure its similarity to another recording of an unknown person, an LLR of -10 means it is e^{-10} times—that is, approximately 22,000 times—more probable of observing the given similarity of the two recordings if they both come from Peter than if one comes from Peter and the other from another person. And conversely, a LLR of 10 means it is approximately 22,000 times more probable of observing the given similarity of the two recordings if one comes from Peter and the other from someone else than if they both come from Peter.

The `SID4-XL` technology contains a submodule called Voice Activity Detection (VAD) that is trained to detect speech and ignore silence and noises. Phonexia recommends that VAD should detect at least 3 seconds of speech in a recording so that it can be used for speaker recognition.

3.1.4 SpeechBrain `spkrec-ecapa-voxceleb`

SpeechBrain is a general-purpose speech toolkit that provides ready-to-use neural network models and Python libraries to perform all sorts of speech analyses (Ravanelli et al. 2021). `spkrec-ecapa-voxceleb` (or simply `spkrec`) is a pre-trained speaker verification model trained on the Voxceleb datasets (Nagrani, Chung, and Zisserman 2017; Chung, Nagrani, and Zisserman 2018). The dataset is composed of data automatically downloaded from YouTube videos.

The SpeechBrain speaker recognizer uses an ECAPA-TDNN model (Desplanques, Thienpondt, and Demuynck 2020) to create speaker "embeddings", i.e., vectors that capture the characteristics of the speaker's voice. The website where the model is published states that in `spkrec` "Speaker Verification is performed using cosine distance between speaker embeddings" (spkrec-ecapa-voxceleb 2021). However, the library code itself shows that in fact the scores are cosine *similarity*, which can have values from -1 (most dissimilar) to 1 (most similar), and the default threshold for deciding if the speakers are the same or different is 0.25.

Even if `SID4-XL4` and `spkrec` have a different range of possible score values, the EER metric can be used for both systems with the same validity because it does not evaluate the *magnitude* of how correct or incorrect a given score is. It only compares scores to possible threshold values, so the two kinds of output, LLR and cosine similarity, can be viewed as two kinds of score scaling.

EER and EER threshold values are commonly reported in evaluations of SID systems; they are not the only metrics for evaluating ASR models, though. The log-likelihood-ratio cost (Cllr) has the advantage of being a "single value summary of system performance" (Morrison and Enzinger 2016), but as the term suggests, it is only suitable for likelihood ratio values, not for other types of ASR scores.[12]

3.2 Language identification

Language identification technologies (LID) can be used to detect the language of an audio recording and are an important component in various applications. Based on the results of LID, phone calls in contact centers can be routed to specific agents who can speak the language; the result of LID can activate a specific speech transcription model in automatic processing pipelines. The incorrect identification of a language in an audio recording can therefore waste significant human and material resources that are needed elsewhere.

LID technologies are typically trained to recognize a large number of languages. When applied to a recording, the technology computes for each language the probability that it is spoken in the recording. LID technologies customarily return only the best-matching language but the probabilities for all other languages can be easily obtained. The accuracy of the system can be easily measured with a data set and language labels for the recordings in the data set. Identification accuracy (ACC) can be defined as the ratio between the number of correctly identified languages and the total number of recordings in the evaluation data set:

$$ACC = \frac{N\ correct}{N\ total} \tag{3.2}$$

3.2.1 Phonexia `LID-L4`

The language identification technology Phonexia `LID-L4` uses the same x-vector architecture as `SID4`, but instead of recognizing speakers it is trained to ignore speaker-specific characteristics and to identify features that are typical for individual languages (Michal Klčo, personal communication). By default, the model distinguishes between 63 languages but it has tools for creating custom sets of languages for identification.

When `LID-L4` processes a recording, it returns a score for each language it knows in the form of a log-likelihood, which means that the score is the natural logarithm of the likelihood that the given language is spoken in the recording. Likelihood values can range between 0.0 (impossibility) and 1.0 (certainty). The log-likelihood is thus

12. The Log Likelihood Ratio Cost metric applies penalties for each wrongly identified pair of recordings. The more the ASR score deviates from the correct decision, the higher the penalty. The metric expects the score to be a likelihood ratio and would produce uncomparable results with another type of score, such as cosine similarity. For a precise definition of Cllr see Drygajlo et al. (2015, 26)

always a negative[13] number—the closer to 0, the higher the score. The likelihoods for individual languages always add up to 1.

3.2.2 SpeechBrain `lang-id-commonlanguage_ecapa`

The language identifier provided in the SpeechBrain toolkit uses the same ECAPA architecture (Desplanques, Thienpondt, and Demuynck 2020) as the SpeechBrain `spkrec` model. `lang-id-commonlanguage_ecapa` (or `lang-id` for short) is based on the CommonLanguage data set, which contains around one hour of audio recordings per language (Sinisetty et al. 2021) and is trained to recognize 45 different languages (lang-id-commonlanguage_ecapa 2021). According to the documentation, the score returned by `lang-id-commonlanguage_ecapa` is a log-posterior: values can be negative (greater than -1) as well as positive (smaller than 1)—the closer to 1, the higher the score.

13. Theoretically, the score could be zero if the system were 100 % sure about the language, which practically never happens because there is always at least a fraction of the likelihood that it can be a different language.

4 Research questions related to foreign accent rating

In this chapter, I look primarily into two topics: Section 4.1 is about how the focus on either accent strength or rater confidence influences accent ratings[14], and Section 4.2 investigates how accent ratings differ in clean recordings and in adverse listening conditions. In order to address these questions, I created several versions of an online accent rating experiment that allowed me to also deal with three supplementary questions: in Section 4.3 I analyze the relationship between articulation rate and foreign accent ratings, in Section 4.4 I look into rater familiarity with talker accent, and finally in Section 4.5 I deal with the topic of native language (mis)match between rater and talker. The point of all the questions is to better understand foreign accent variation in the data set that will be used in an evaluation of biometric speech technologies. The question of how a foreign accent influences speech technologies' performance will be the topic of Chapter 5. In these two chapters, I will report a number of statistical tests performed by Dr. Richard Andrášik in R (R Core Team 2021), who also contributed most of the descriptions of the statistical methods used.

In all versions of the experiment I tried to avoid the frame-of-reference issue explained in Section 2.3.1. The problem lies in the observation that the definition of scale endpoints in terms of either native language or foreign accent makes raters activate different mental concepts that are not uniform across raters, and causes a difference in the scaling of accent ratings. In order to balance these two possibilities of defining the frame of reference, the endpoints of the rating scale in my experiments always contained reference to both foreign accent and nativeness (see Section 6.2.5 for more details).

4.1 Formulation of the foreign accent rating task

Previous research summarized in Section 2.3 shows that foreign accent ratings are influenced by the formulation of the rating task. Typically, raters are asked to rate

14. A logical extension to the options of asking for accent strength and for rater confidence would be to ask listeners to both rate the perceived accent strength and also evaluate their confidence in the rating. This would have made a nice third version of the experiment, only it would have required significantly rewriting the experiment script in PsyToolkit (see Section 6.2.5), and, more importantly, it would have required increasing the total number of raters (see Section 6.2.6), which turned out to be rather unrealistic. In addition, asking listeners to rate two aspects of each sample would have prolonged the already exhausting experiment beyond bearable limits, or else the number of samples per rater would have to be lower, requiring even more raters.

the degree of foreign accent, sometimes also to express their confidence about the rating. But can the two tasks be used interchangeably? Or does one of them help raters to better tell apart native and non-native talkers? Based on the literature overview and on my personal experience with foreign accents, I would suggest the following hypothesis:

Hypothesis 4.1.1: *People can notice even a weak foreign accent, and the task of rating their* degree of confidence *that a speaker has a foreign accent, as opposed to being a native speaker, can help them to distinguish non-native talkers from native talkers more clearly than if their task is to rate the* strength *of the speaker's* foreign accent*. Native speakers—by definition those without a foreign accent—should not be affected by the rating task.*[15]

If Hypothesis 4.1.1 is correct, and we calculate average ratings from the two groups of raters, then we can expect the following results:

Prediction 4.1.1: *If we look at all talkers together, the absolute values*[16] *of the ratings will cluster more tightly in the higher* Score *region for the group who rate their confidence, while the absolute values of* Score *from raters who focused on accent strength will be more evenly distributed along the whole scale and the median will be smaller.*

Prediction 4.1.2: *The listeners who rate their confidence that talkers have a foreign accent will on average rate non-native talkers significantly closer to the "foreign accent" end point of the scale than the raters who focus on the degree of foreign accent.*

Prediction 4.1.3: *Prediction 3: Ratings for native talkers will on average be closer to the "native speaker" end point of the scale, and there will be no significant difference between mean ratings of the listeners who rate their confidence and the listeners who rate the strength of the foreign accent.*

We will refer to the raters who focus on the strength of the perceived foreign accent as the "How strong" group, and the raters who focus on their confidence about the presence of a foreign accent will be referred to as the "How sure" group.

4.1.1 Method

The way foreign accent ratings were obtained is described in detail in Section 6.2.5; here I summarize only the salient points: The same set of stimuli was presented to

15. Admittedly, this is a somewhat controversial claim since—as has been shown in numerous studies—even some native speakers receive accent ratings similar to those of proficient non-native talkers (Bongaerts, Planken, and Schils 1995).

16. As explained in Section 6.2.5, the ratings could go from -300 ("native speaker") to 300 ("foreign accent"). When looking at native and non-native speakers together, we expected a bimodal distribution of the *Score*. Using the absolute values helped us to see the extreme values.

two randomly[17] selected groups of raters. One group had to answer the question "How sure are you that the speaker has a foreign accent or that the person is a native speaker of English?", and each member rated each sample on a continuous scale with the end points labeled as "definitely has foreign accent" and "definitely is native speaker"[18]. The other group had to answer the question "How strong is the foreign accent of the speaker or how much does the person sound like a native speaker of English?", and each member rated each sample on the same continuous scale with the end points labeled as "very strong foreign accent" and "sounds like a native speaker".

A continuous scale was chosen for the rating rather than the usual Likert scale in order to avoid the problems connected with the statistical analyses of ordinal data produced by the Likert scale (see Section 2.3.3). If we want to identify whether a continuous random variable is related to a discrete random variable, it is necessary to separate the records into groups according to the discrete variable and subsequently compare the probability distribution of the continuous variable across the groups. In our case, the continuous random variable is the foreign-accent score, so we will refer to the variable simply as *Score*. The discrete random variable, on the other hand, is the *Rating question* that defines only two groups: How strong and How sure. With only two groups defined by the discrete variable, we can use the Kolmogorov-Smirnov test (KS test; Corder and Foreman 2009). This test compares the empirical Cumulative Distribution Functions (CDF). The null hypothesis about their equality is rejected when the distance between them is greater than a certain threshold.

Results achieved by the KS test can be complemented by the two-sample Wilcoxon test which compares the medians of the groups (ibid.). Both the KS test and the Wilcoxon test are non-parametric tests (i.e., there is no assumption on the data distribution and there is no need to verify the normality of the data). The statistical results are illustrated by graphs of the empirical CDFs and boxplots below. The KS test and the Wilcoxon test were performed in R with the routines `ks.test` and `wilcox.test`, respectively.

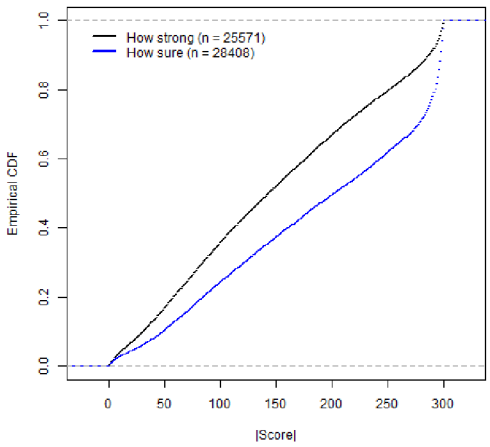
4.1.2 Results

In Prediction 4.1.1—absolute values of all ratings will be higher for the "How sure" group and more evenly distributed for the "How strong" group—we considered native and non-native talkers together (n = 42). We were able to use the absolute value of *Score* to see the extreme values even if the two groups had their median values on the opposite sides of the scale. Both the KS test and the Wilcoxon test rejected the null

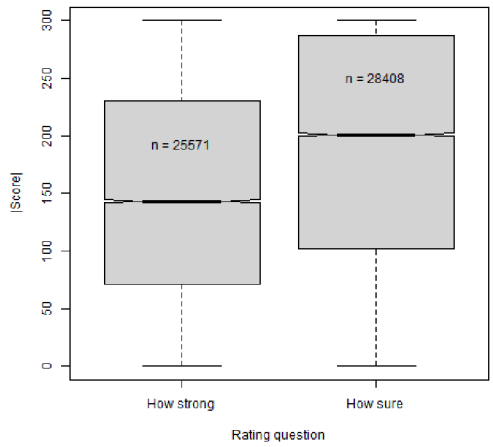
17. Random selection of participants in this type of research is, of course, an illusion. If the raters are supposed to be representative of, let's say, the whole population of native speakers of English, the method of selection of participants would make sure that every single native speaker of English had the same probability of being selected for the sample. In reality, it can turn out to be so difficult to get anybody to do such an experiment that a researcher cannot afford being too picky when it comes to participants. So when I claim that raters in my experiment were randomly selected, what I mean in fact is that participants were randomly assigned a different version of the experiment (see Section 6.2.5 for more details).

18. As one participant pointed out in the feedback section, this should have read "definitely is a native speaker". Hopefully most participants were not too picky and did not get too distracted by the mistake.

hypothesis (both p-values were lower than 0.0001). Hence, the probability distribution of $|Score|$ in groups "How strong" and "How sure" was different. Furthermore, it can be seen from Figure 1 that $|Score|$ in the group "How sure" tended to be higher than $|Score|$ in the group "How strong". The difference between medians amounted to 59.



(A) The probability distributions of $|Score|$ for the "How strong" and "How sure" groups. Values on the vertical axis show what percentage of $|Score|$ in both groups were below a certain $|Score|$ value. For illustration, while more than 50 % of $|Scores|$ in the group "How strong" are below 150, in the group "How sure" it is less than 40 % of $|Scores|$.



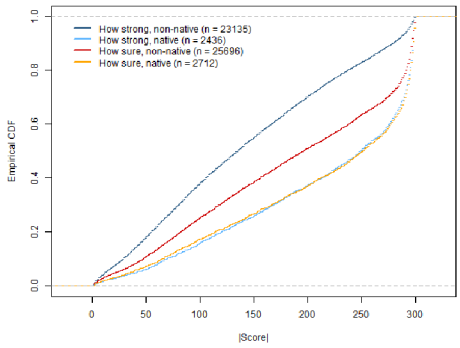
(B) The difference in accent $|Score|$ between the "How strong" and "How sure" groups. The boxes, representing the Interquartile Range, and the medians, represented by the black horizontal lines inside the boxes, show that $|Score|$ tended to be higher in the group "How sure", whose members rated their confidence that they heard a foreign accent or that the talker is a native speaker.

FIGURE 1: Absolute values of *Score* for two versions of the *Rating question* for native and non-native talkers together (n = 42). The n values in the boxplots and in the CDFs show the total number of observations in the samples.

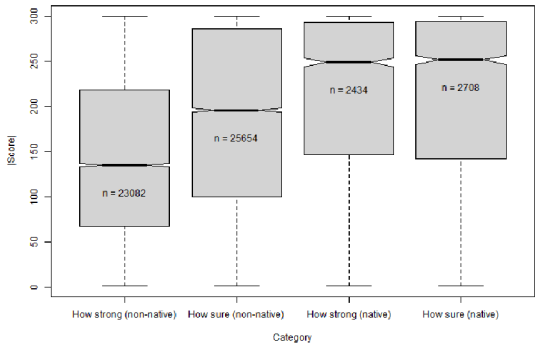
When we look at native and non-native talkers separately in Figure 2, we can see that they behaved differently from each other. While non-native talkers received a generally lower $|Score|$ from people who rated "How strong" than from people who rated "How sure", this was not true for native talkers for whom the absolute values of *Score* apparently did not differ between both rater groups.

In Prediction 4.1.2—the "How sure" group will rate non-native talkers significantly closer to the "foreign accent" end point than the "How strong" group—we only considered non-native talkers. Both the KS test and the Wilcoxon test rejected the null hypothesis (both p-values were lower than 0.0001). Hence, the probability distribution of *Score* in groups "How strong" and "How sure" was different. Furthermore, it can be seen from Figure 3 that *Score* in the group "How sure" tended to be higher (closer to 300) than *Score* in the group "How strong". The difference of medians amounted to 76.

In Prediction 4.1.3—ratings for native talkers will be closer to the "native speaker" end point in both groups, "How sure" and "How strong"—we only considered native talkers to corroborate results from Prediction 4.1.1 for which we measured the absolute values of *Score*. The results are shown in Figure 4. The KS test rejected

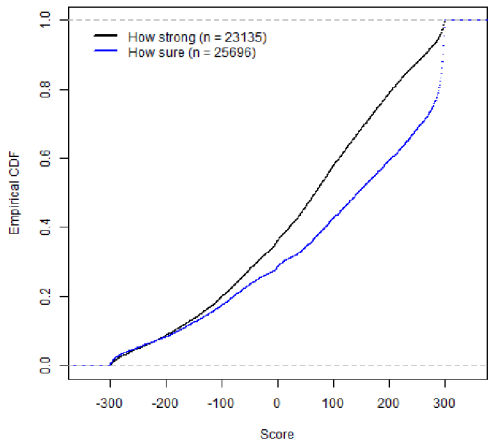


(A) Probability distributions broken down for native and non-native talkers and for two versions of the *Rating question*. We can see that the group "How strong" gave in general a less extreme $|Score|$ than the "How sure" group, however, only for non-native talkers, and native talkers received a generally higher $|Score|$.

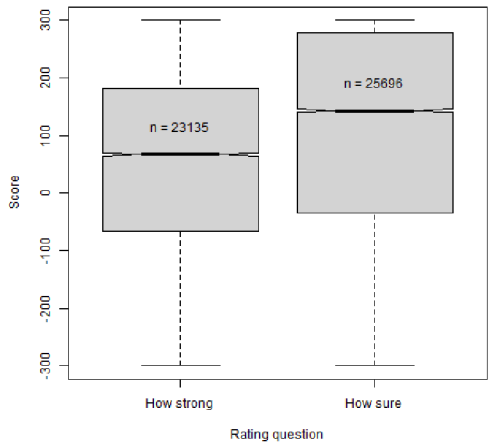


(B) The difference in accent $|Score|$ for two versions of the *Rating question* broken up for native and non-native talkers. Listeners rating their confidence that they heard a foreign accent gave a more extreme $|Score|$ than listeners rating the strength of a perceived accent in the case of non-native talkers. In the case of native talkers there was apparently no difference in $|Score|$ between the "How strong" and "How sure" rater groups.

FIGURE 2: Absolute values of *Score* for two versions of the *Rating question* separately for native (n = 4) and non-native (n = 38) talkers.



(A) Probability distributions of *Score* for the "How strong" and "How sure" groups. The distributions differ mainly in the positive *Score* region, which is the region of "strong foreign accent" and "high confidence of foreign accent", but they are almost identical in the negative *Score* region of "native speaker".

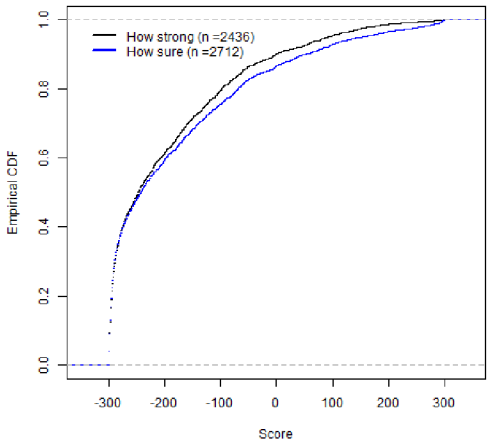


(B) The difference in accent *Score* for the "How strong" and "How sure" groups. Listeners rating their confidence that they heard a foreign accent ("How sure") gave overall higher *Score* than listeners rating the strength of a perceived accent.

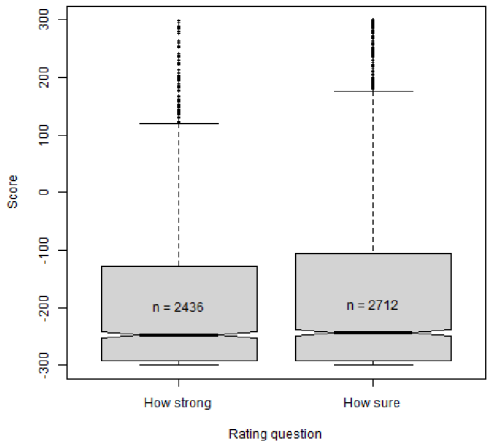
FIGURE 3: Two versions of the *Rating question* in the case of non-native talkers (n = 38).

the null hypothesis (p-value = 0.0144) while the Wilcoxon test did not reject it (p-value = 0.0867). It seems that there was only a slight difference in the probability distributions. Concerning the medians, no statistically significant difference

was found. Their difference was only 5. From a practical point of view, the groups can be considered almost the same.



(A) Probability distributions of *Score* for two versions of the *Rating question* in the case of native talkers. The concave shape of the closely correlated curves supports the idea that native talkers were almost identically recognized as such by both rater groups.



(B) The overlapping notches in the boxplots illustrate the negligible difference in the *Score* for the two *Rating questions* with native talkers.

FIGURE 4: Two versions of the *Rating question* in the case of the four native talkers.

4.1.3 Discussion

In this section, we looked at how the formulation of the rating task, or the *Rating question*, influenced foreign accent ratings for native and non-native English talkers. Previous studies have found that the way the rating task is formulated can determine the "frame of reference" of the raters. In other words, by specifying whether listeners should rate either foreign accent or "nativeness", raters are led to activate their implicit or explicit notions of pronunciation standards, which can be more or less homogeneous or may be missing altogether. We tried to avoid this pitfall by always including in the rating task a reference to both nativeness and foreign accents, however, the decision not to inform raters about the composition of the data set[19] may have caused other problems. From the feedback provided by the raters, it seemed that at least some of them struggled with the rating task, which forced them to place a rating on the scale between "native speaker" and "has foreign accent", when they would have rather chosen between "is or is not native speaker" or "has or does not have foreign accent". In any case, the decision not to inform participants about the presence and number of native speakers in the data set, or about the regional accents they would encounter, seemed to be the right one with respect to making the results more realistic by not giving the raters any advantage of this sort.

19. This decision was motivated by the effort not to prolong the accent-rating experiment beyond what seemed necessary.

The question we wanted to address in this section was how the focus on either the strength of accent or on the rater's confidence influenced foreign accent ratings. In other studies, these two measures have been used in combination, i.e., listeners were first asked to make a judgment about nativeness or non-nativeness of the speaker and then to give a confidence rating of their choice (e.g., Ulbrich and Mennen 2015, Schmid and Hopp 2014). This procedure certainly makes rating experiments last longer, hence more demanding for raters, and it is questionable whether it provides any information that cannot be gained from a single rating. If confidence and accent-strength ratings turned out to provide comparable results, raters should be spared of the extra chore, especially considering that the two ratings are generally combined into one single measure (more about this procedure in Section 4.4.3).

In our data set, however, we expected raters to react differently to the task of rating their confidence of a perceived foreign accent, and of rating the accent per se, because we hypothesized that people can be confident even about the perception of a weak accent in non-native speakers. In Section 4.1.2 we presented the results that seemed to confirm the predictions—non-native talkers received more average ratings from the "How strong" group of raters than from the listeners who focused on their confidence of the perceived accent. There was no difference in the ratings for the native speakers—it appeared that raters were able to identify native speakers equally well when they focused either on their confidence or the accent itself. It is important to note, however, that there were only four native speakers in the data set, as opposed to 38 non-native speakers.

Even if, in the case of non-native talkers, the ratings from the two groups were statistically different—i.e., the "How sure" group had a higher median and the empirical CDFs were different too—the two Scores were still strongly correlated. Figure 5 shows the correlations between the median values of Score for individual talkers and for individual recordings.

The boxplots and scatter plots with median values do not show one important aspect of the *Score* for non-native talkers, which is that at least some raters used the continuous scale as a binary or ternary one. This becomes apparent when we put all individual ratings into a histogram (see Figure 6 for more details).

4.2 Foreign accent rating in adverse listening conditions

Previous research mentioned in Section 2.7 shows that foreign accent perception in noisy data is more difficult for non-native raters. The difference from native raters is especially noticeable if an accent rating is performed on larger linguistic units such as words and sentences (Lecumberri, Cooke, and Cutler 2010). The majority of adverse-listening-conditions research has focused on adding noise, sometimes also on reverberation, so the question of how foreign-accent ratings would be influenced by low acoustic quality in simulated telephone speech—characterized primarily by filtering out frequencies rather than adding noise—does not have an obvious answer.

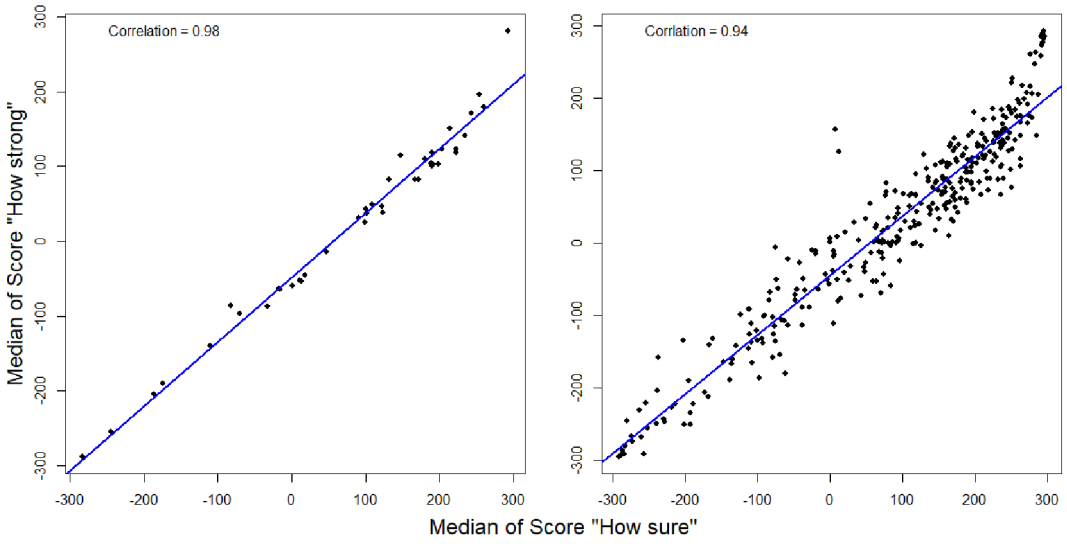


FIGURE 5: Correlation between "How strong" and "How sure" scores. The plots show a slightly stronger correlation between median scores for individual talkers than for individual recordings. The outliers in the upper-right corner mostly correspond to the Ukrainian/Russian talker who was not only rated with high confidence as foreign-accented, but also as having an above-average accent strength. Two interesting outlier recordings have been rated as rather strongly accented (*Score* 125 and 156) but most raters were not sure whether the talker is native or non-native.

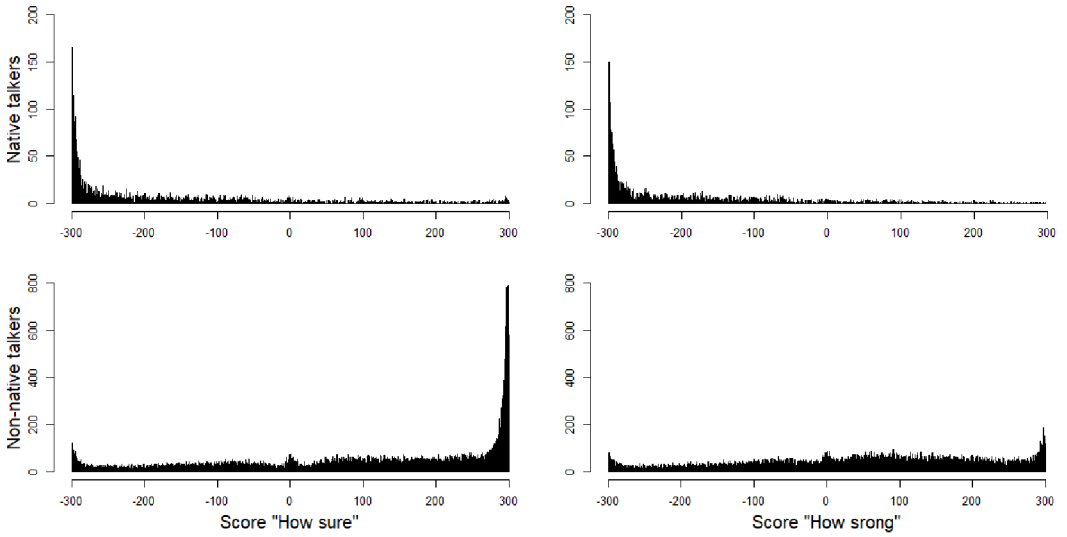


FIGURE 6: The highly non-normal distribution of both the "How sure" and "How strong" *Score* values in the case of non-native talkers shows that at least some raters used the continuous scale as a ternary one: the spikes at the -300 and 300 edges, and a spike at *Score* 0—more prominent in the "How sure" group—could be understood as "no", "yes", and "I don't know" answers to the question: "Does the speaker have a foreign accent?" It is possible that the dips surrounding the middle spike were influenced by the design of the rating interface, which contained the "replay" and "next" buttons that could have been used as visual cues for choosing the *Score* (see Section 6.2.5). The histograms for native talker ratings uses a different scale, the only obvious spike is the one at the "Native speaker" end of the scale.

Some studies suggest, however, that frequency filtering similar to that of simulating phone call quality reduces raters' ability to distinguish different degrees of foreign accent (Volín and Skarnitzl 2010), so we can hypothesize the following:

Hypothesis 4.2.1: *Suprasegmental features, such as articulation rate, can interact with segmental features to trigger high foreign accent ratings in high-quality recordings of non-native speech, but in filtered recordings, segmental features are degraded and their contribution to the perception of a strong foreign accent is reduced, even if temporal suprasegmental features, such as articulation rate, are largely unaffected. This degradation also makes it more difficult to differentiate native speakers from non-native speakers.*

4.2.1 Method

Two versions of the stimuli set were presented to two randomly selected groups of raters. One version consisted of original high-quality recordings (in the following referred to as "Original"), and the other consisted of the original recordings processed to simulate low-quality landline-telephone characteristics, in the following referred to as "Phone" (see Section 6.2.4). The raters had to answer one of the questions described in the previous section: "How sure are you that the speaker has a foreign accent or that the person is a native speaker of English?" or "How strong is the foreign accent of the speaker or how much does the person sound like a native speaker of English?" (see Section 4.1.1 for more details).

The difference in probability distributions and medians between samples was measured like in Section 4.1 with the KS test and Wilcoxon test, respectively. The difference in variability of *Score* in two sub-samples was measured as the difference in the interquartile range of the subsets and it was performed with Levene's test (Derrick et al. 2018) using the R routine `levene.test`.

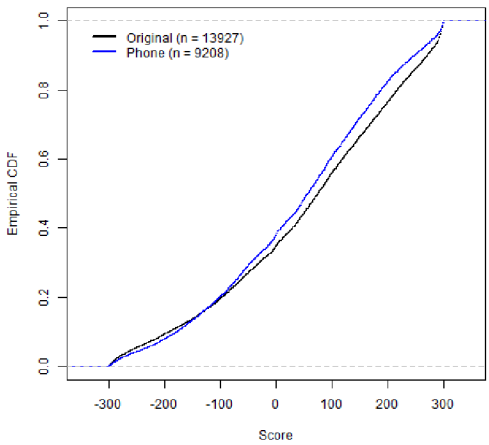
In Section 4.1 we found that native talkers received lower accentedness ratings than non-native talkers, and also that in the case of non-native talkers, different lower ratings were elicited for the "How strong" question than for the "How sure" question. For these four groups, the following predictions were tested separately:

Prediction 4.2.1: *Raters will be better able to tell apart native and non-native talkers and distinguish various degrees of foreign accent in the non-native talkers based on the "Original" recordings, but not in the "Phone" recordings. In other words, in the case of non-native talkers, there will be a significantly larger variability of* Score *values in the "Original" recordings, reflecting the real variability of accents, and the* Score *values will be on average significantly higher than in the "Phone" recordings, because more non-native features will be distinguishable in the "Originals".*

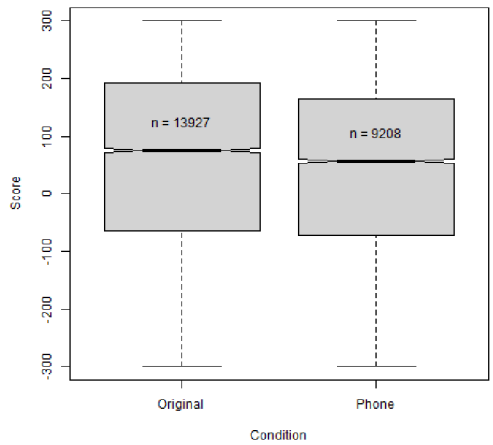
Prediction 4.2.2: *In the case of native talkers, there will be a significantly larger variability of* Score *values in the "Phone" recordings, reflecting the increased difficulty to tell native talkers apart from non-native talkers. The* Score *values will be significantly higher (i.e., farther away from the "native speaker" end point) than in the "Original" recordings.*

4.2.2 Results

Let us begin with the results for non-native talkers. First, only the group "How strong" was analyzed. The results are shown in Figure 7. Both the KS test and the Wilcoxon test rejected the null hypothesis (both p-values were lower than 0.0001). Hence, the probability distribution of *Score* in the "Original" and "Phone" groups was different. The difference between medians amounted to 19. In general, the "Original" recordings received higher *Scores*, i.e., were rated as more foreign-accented. Variability was significantly lower in the "Phone" group than in the "Original" group (p-value < 0.0001).



(A) The probability distributions of *Score* show that the "Original" recordings received a higher overall *Score* than the "Phone" recordings from the "How strong" group.

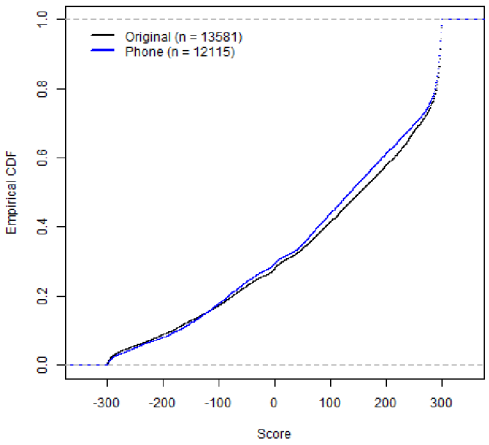


(B) "Original" recordings had a higher median and a larger variability of *Score*, which can be seen as a bigger box that stands for the interquartile range.

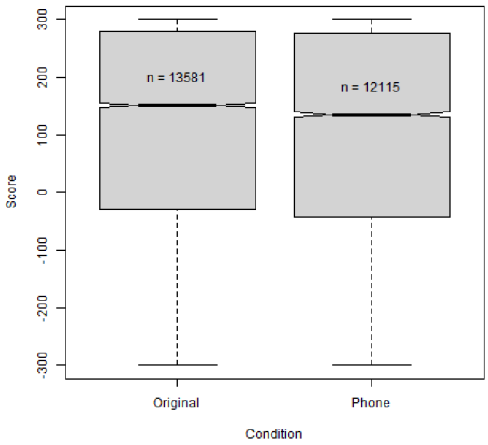
FIGURE 7: Two versions of the data set: "Original" and "Phone" in the case of non-native talkers for the "How strong" question.

Second, the group "How sure" in the case of non-native talkers was evaluated. The results are shown in Figure 8. With respect to the probability distribution of *Score*, the results pointed in the same direction as for the "How strong" group (compare Figures 7 and 8), which means that the "Original" recordings received a higher *Score* on average. The difference between medians amounted to 17. However, the difference in variability of *Score* in groups "Original" and "Phone" was not statistically significant.

The same analysis was performed also for the native talkers. First, the group "How strong" was evaluated. The results are shown in Figure 9. Both the KS test and the Wilcoxon test rejected the null hypothesis (both p-values were lower than 0.0001). Hence, the probability distribution of *Score* in the "Original" and "Phone" groups was different; the "Phone" recordings received on average a higher *Score* than the "Original" recordings. The difference between medians was 62.5. The variability of *Score* was significantly higher in the "Phone" group (p-value < 0.0001).

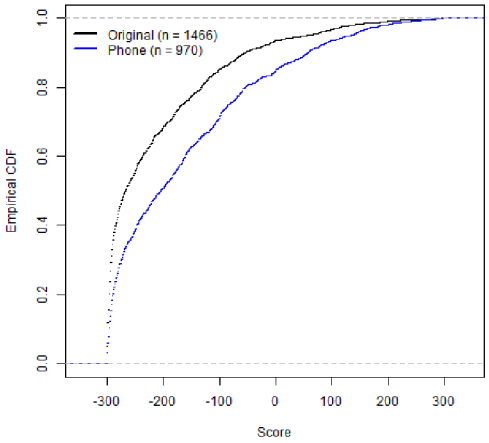


(A) The probability distributions of *Score* for the "Original" and "Phone" recordings. The "Original" recordings received a significantly higher *Score*, however, it can be seen that the two probability distributions follow very similar paths.

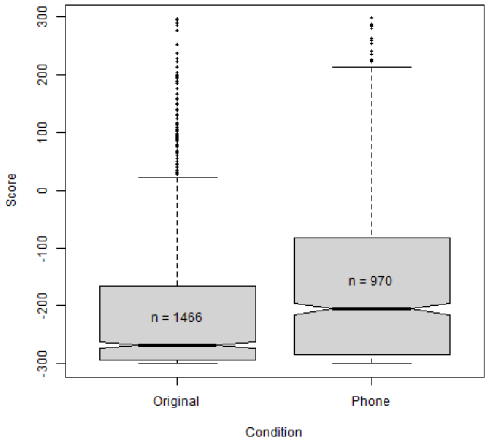


(B) The difference in accent *Score* between the "Original" and "Phone" recordings. The "Original" recordings have a significantly higher median, even though the difference is not a large one. The variability of *Score* is not significantly different.

FIGURE 8: Two versions of the data set: "Original" and "Phone" in the case of non-native talkers for the question "How sure".



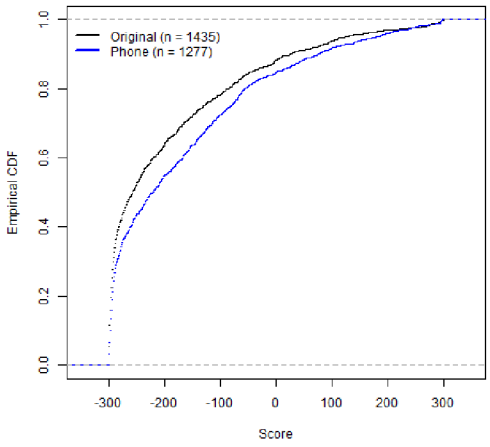
(A) The probability distributions of *Score* showing that the "Phone" recordings received clearly higher *Score* than the "Original" recordings.



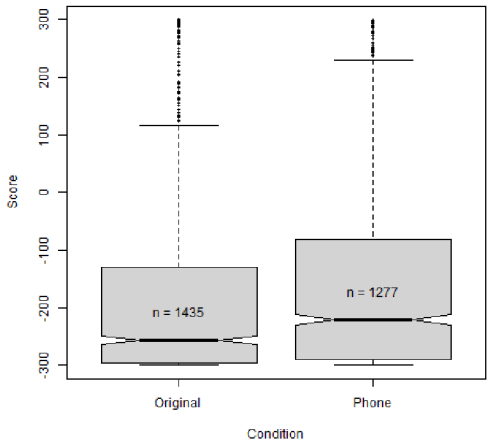
(B) The "Phone" recordings have a significantly higher median and greater variability of *Score* values.

FIGURE 9: Two versions of the data set: "Original" and "Phone" in the case of native talkers for the question "How strong".

Second, the group "How sure" for native talkers was evaluated. The results, which can be seen in Figure 10, were the same as for the group "How strong"—the "Phone" recordings received a higher *Score* on average than the "Original" recordings. The difference between medians was 36. The variability of *Score* was significantly higher in the "Phone" group (p-value = 0.0003).



(A) The probability distributions of *Score* showing that just like in the case of the "How strong" group, the "Phone" recordings received clearly higher *Score* than the "Original" recordings.



(B) The "Phone" recordings have a significantly higher median of and greater variability of *Score* values even though the difference is not as pronounced as in the "How strong" group.

FIGURE 10: Two versions of the data set: "Original" and "Phone" in the case of native talkers for the question "How sure".

4.2.3 Discussion

In this section, we analyzed the relationship between foreign-accent ratings and the listening conditions in the stimuli. In confirmation of our predictions, we found that the adverse listening conditions of the simulated telephone recordings made it more difficult for raters to identify foreign accent in the speech of non-native talkers and, at the same time, native talkers were rated as more strongly accented in the "Phone" recordings. The largest difference between the "Original" and "Phone" recordings was observed for native talkers in the case of the rating question "How strong". This confirms the finding of many previous studies that even native speakers can be perceived as having a mild foreign accent under certain conditions.

As for the interpretation of the results for the non-native talkers, the statistics that collapsed all non-native talkers into one group covered up an interesting fact that became obvious when we looked at the medians of *Score* for individual talkers in the "Phone" and "Original" recordings separately. As can be seen in Figure 11, not all non-native talkers were rated as less accented in the "Phone" recordings. In fact, most of the non-native talkers who received in general more native-like ratings were rated as less accented in the "Original" recordings, just as the native talkers. Table 4.1 shows another interesting aspect of this—while for native talkers the difference in the *Score* median between the lowest- and highest-scoring talker was smaller in the "Original" recordings, it was the other way round for the non-native talkers, because the non-native "Low scorer" was rated in a way similar to the native talkers, i.e., she too received lower accentedness scores in the "Original" recordings.

When we divided the talkers into two groups based on whether the median of all the *Scores* they received in both listening conditions was smaller or greater than 0—the hypothetical point of rater indecision—the two groups behaved in a similar way as the original groups divided by native language: the "Low Scoring" group received significantly higher accent *Scores* in the "Phone" recordings than in the "Originals" (p-values < 0.0001) and, vice versa, the "High Scoring" group received significantly higher *Scores* in the "Original" recordings (p-values < 0.0001).

Talker group – Rec. type	"Low scorer"	"High scorer"	Difference
Native – Original	−292	−197.5	94.5
Native – Phone	−275	−151	124
Non-native – Original	−134	292	426
Non-native – Phone	−111	290.5	401.5

TABLE 4.1: Median values for the lowest- and highest-scoring talker in each talker group show that the difference between the native low- and high-scorers is bigger in the "Phone" recordings, whereas with non-native talkers it is bigger in the "Original" recordings.

In conclusion, the adverse listening conditions in the "Phone" recordings were linked to more average foreign accent scores when it came to both ratings of accent strength and ratings of rater confidence in perceiving a foreign accent. This was found for native speakers who received higher *Scores* in the "Phone" recordings, as well as for non-native speakers who received lower or higher *Scores* in the "Phone" recordings, depending on whether they were in general rated as more or less foreign-accented, respectively.

4.3 Articulation rate and foreign accent rating

In earlier research, articulation rate has been found to negatively correlate with foreign accent ratings. Volín and Skarnitzl (2010) found that faster talkers received better accent scores, i.e., they sounded more native-like to the raters. The study also found that the correlation was weaker in filtered speech, apparently because talkers with a strong foreign accent received lower accent scores in adverse listening conditions. In order to find out if articulation rate had any influence on accent ratings in the data set—especially whether it played any role in accent ratings in adverse listening conditions—first, a baseline measurement was performed that compared the articulation rates between native and non-native talkers.

Prediction 4.3.1: *The four native talkers will have a higher articulation rate than the 38 non-native talkers.*

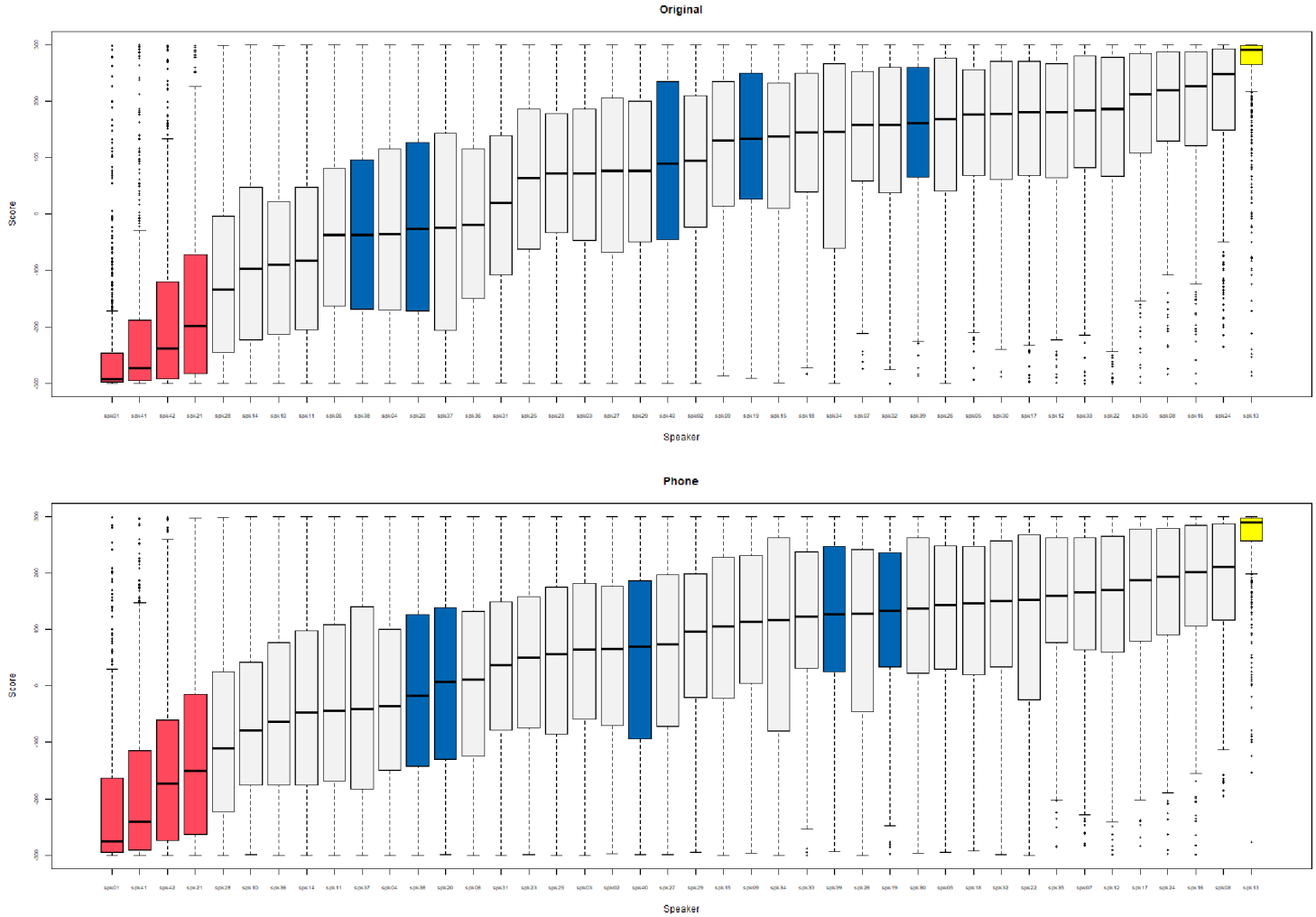


FIGURE 11: Medians of *Score* ("How sure" and "How strong" ratings together) per talker in the "Original" and "Phone" recordings. Native English speakers are shown in red, Slovak speakers in blue, the Ukrainian/Russian speaker in yellow, and the rest are native speakers of Czech.

4.3.1 Method

Articulation rate was measured semi-automatically with the help of the Praat script `syllable-nuclei-v2`[20] (De Jong and Wempe 2009). The script measures articulation rate as the ratio between an automatically-detected number of syllables and phonation time (that is the time of speaking from which all pauses and possibly also voiceless sounds—if they take too much time and fall into the category of pauses—are excluded). However, the script's automatic syllable detection does not always pick up unstressed and reduced syllables—exactly the kind of syllables that are interesting in an accent-rating experiment because Czech and Slovak talkers, as opposed to native talkers, can be expected not to reduce unstressed syllables. The script could therefore inaccurately present native talkers as having slower articulation because it might skip more of their unstressed syllables than it would with the non-native talkers. Since the number of syllables in the canonical pronunciation of the "The North Wind and the Sun" samples was known, articulation rate was alternatively measured as the number of ideal syllables divided by the phonation time as measured by the `syllable-nuclei-v2` script. This is in line with articulation rate measurements reported by Volín and Skarnitzl (2010, 1013) who also counted the number of syllables according to the dictionary forms of the words.

Articulation rate was only measured for the original recordings in the FA rating experiment, not for the phone simulations—based on the assumption that the speech rate remains the same in the phone recordings—because the temporal characteristics are largely unchanged by the phone simulation, except maybe for reducing the voicing in the final sounds at the rightmost phrase boundaries, which in some recordings contained low-intensity creaky voice. The annotation of silences—automatically performed by the script—was checked manually for each stimulus because sometimes the script did not identify an initial or final silence correctly, which would have made the measurement of phonation time inaccurate.

The correlation between *articulation rate* and FA ratings was measured with Spearman correlation coefficients. The coefficients can have values from -1 (very strong inverse correlation) to 1 (very strong positive correlation). A coefficient value of 0 means there is no correlation.

4.3.2 Results: native vs non-native talkers

The random continuous variable *Articulation rate* in the two talker groups was compared by the difference of medians and by the difference of probability distribution functions. The difference in medians was statistically significant (p-value = 0.0084; difference = 0.6 syll./s). The difference in the second measure, which describes the whole distribution of probabilities of the *Articulation rate* values, was not statistically significant. The probability distribution test is conservative, which means that the

20. The script that was actually used, https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2, is a modified version of the original (De Jong and Wempe 2009) script.

difference has to be rather large in order to turn out statistically significant. Based on the difference of medians, therefore, it seems that there really is a difference in *Articulation rate* of the native and non-native talkers in the data set, but the difference is not a large one. Besides, the number of samples considered in this analysis was rather small, especially in the case of the native talkers, which can be illustrated by the large confidence intervals of the medians in Figure 12.

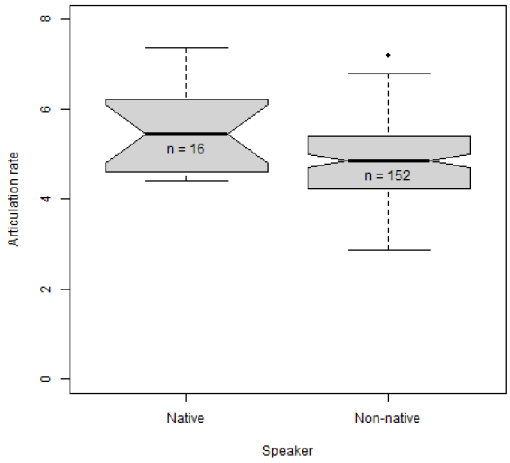


FIGURE 12: The difference in the *Articulation rate* of four native and 38 non-native talkers. The medians are significantly different, however, the probability distribution functions are not. Furthermore, the overlapping notches in the boxplots suggest that even the difference between medians is not very conclusive.

4.3.3 Results: correlation with FAR

We further tested two predictions regarding the question whether there is a correlation between *Articulation rate* and accent *Score*. The correlation was measured separately for native and non-native talkers because there seemed to be a difference in *Articulation rate* between the groups based on the results shown in Figure 12, and, more importantly, the groups differed in the way the adverse listening conditions affected their accent ratings. Recall that while native talkers were rated as significantly more accented in adverse listening conditions than in the "Original" recordings, for non-native talkers accent ratings were lower in the "Phone" recordings—the difference was statistically significant, but rather small in the case of the non-native speakers (see Section 4.2.2). We did not expect, therefore, a similar effect as was reported in (Volín and Skarnitzl 2010), i.e., that more adverse listening conditions would weaken the correlation between *Articulation rate* and accent Score. Instead, the "Original" and "Phone" recordings were treated separately based on the following hypothesis:

Hypothesis 4.3.1: *In the "Phone" recordings, segmental features are degraded and, as a result, temporal features such as articulation rate gain more prominence as a cue for foreign accent rating.*

Prediction 4.3.2: *Higher* Articulation rate *will correlate with lower accent* Score *in native as well as non-native talkers.*

Prediction 4.3.3: *Prediction X.3: The inverse correlation between* Articulation rate *and Score will be stronger in the "Phone" recordings than in the "Original" recordings.*

The results are presented in tables 4.2 and 4.3 as Spearman correlation coefficients. The coefficients were calculated separately for the "How sure" and "How strong" groups since the two groups were shown to behave differently (at least when it came to native talkers) in Section 4.1.

Talker group – Recording type	Spearman correlation coeff.
All together	-0.1902
Native – Original	-0.2248
Native – Phone	-0.2285
Non-native – Original	-0.1028
Non-native – Phone	-0.1445

TABLE 4.2: Spearman correlation coefficients for *Articulation rate* and accent *Score* (question "How sure"). Coefficients closer to -1 signify stronger inverse correlations. It appears that lower *Articulation rate* can be linked to higher accent *Score*, but the relationship is not very strong. It is stronger with native talkers than with non-native talkers. In non-native talkers, it is stronger with the "Phone" recordings than with the "Original" recordings, but the difference between the two is very small.

Talker group – Recording type	Spearman correlation coeff.
All together	-0.2048
Native – Original	-0.2269
Native – Phone	-0.2763
Non-native – Original	-0.1461
Non-native – Phone	-0.1319

TABLE 4.3: Spearman correlation coefficients for *Articulation rate* and accent *Score* (question "How strong"). Coefficients for two talker groups in two recording types. The results are similar to those for the "How sure" question above in that the negative correlation is stronger with native talkers. The difference is that in the case of the "How strong" question, it is the native talkers with whom the correlation is stronger in the "Phone" recordings. In non-native talkers, the "Original" recordings show a stronger relationship with FA than the "Phone" recordings, but the difference is even smaller than with the question "How sure".

4.3.4 Discussion

In this section, we looked at whether *Articulation rate* differed in the native and non-native talkers in our data set and whether it was correlated with the accent ratings for the two talker groups. The baseline prediction that native speakers would in general have a higher *articulation rate* was marginally confirmed. Although the number of observations in the native-speaker group was rather low and did not allow any strong conclusions, the results were in line with previous findings that native speakers tend

to speak at a higher rate than non-native speakers when instructed to read sentences at a comfortable, self-determined speaking rate (Munro and Derwing 1998).

Our next prediction was also marginally confirmed—there was a negative correlation between *articulation rate* and foreign-accent ratings, i.e., faster talkers received lower accentedness scores. However, the correlation was not very strong in native talkers, and in non-native talkers it was even weaker. One possible explanation of the difference between the native and non-native speakers could be that if you are a native speaker, there are probably few segmental features that would make raters believe you are in fact an L2 speaker. Therefore, if you happen to speak slowly, this is what sets you apart from those native speakers who speak faster than you, and probably get lower accentedness scores. When, in contrast, you are a non-native speaker and you speak fast, there may still be segmental features that give you away as a non-native speaker, and if you happen to have a strong foreign accent in terms of segmentals, your fast articulation rate may even contribute to a lower intelligibility of your speech, and possibly even strengthen the perception of your accentedness (see Munro and Derwing 2001).

4.4 Raters' familiarity with talker accents

Based on the literature review in section 2.2, it is not very clear whether familiarity with talker accents will help raters spot the accent more accurately, especially if the accent is weak. Familiarity with native accents will likely influence accent ratings differently from non-native accents. I tried to capture the difference in the following two hypotheses:

Hypothesis 4.4.1: *Familiarity with native accents of English helps raters distinguish native talkers from non-native talkers.*

Hypothesis 4.4.2: *Hypothesis X.2: Good familiarity with a non-native accent in English can be expected predominantly in people who are themselves speakers of that accent. This makes them less sensitive to the accent, and less able to distinguish it from native speech.*

4.4.1 Method

We analyzed the same FA ratings as described in section 4.1. Raters self-reported their familiarity with the English accents included in the stimuli set: British, American, Czech, Slovak, and Ukrainian/Russian English. Further, they could specify their familiarity with any other native or non-native accents in English. FA ratings from five raters who did not specify familiarity with English accents were removed from the analysis.

The FA rating data set was split into three parts based on talker native language: English (referred to as "Native" talkers) and Czech and Slovak (together referred to as "Non-native" talkers). There were four "Native" talkers (two from the

U.S. and two from the U.K.), 32 Czech talkers and five Slovak talkers. The single Ukrainian/Russian talker was not included in the analysis. Spearman correlation coefficients were calculated for the three talker groups and separately for the questions "How sure" and "How strong".

4.4.2 Results

For Hypothesis 4.4.1 we tested the following prediction:

Prediction 4.4.1: *Raters' familiarity with native accents of English will be inversely correlated with the accent* Score *given to native talkers, and will be positively correlated with the accent* Score *given to non-native talkers.*

In Table 4.4 it can be seen that, in the case of the "How sure" question, the correlation was negative for native talkers and positive for non-native talkers, as predicted. The inverse correlation was stronger in the case of the American accent than in the case of the British accent and other native accents. In the case of the "How strong" question (Table 4.5), however, there was no strong correlation, especially in the case of non-native talkers.

Talker group	Raters' accent familiarity	Spearman coeff.
Native	British	-0.1349
Native	American	-0.2677
Native	Other native	-0.1447
Non-native	British	0.2100
Non-native	American	0.2137
Non-native	Other native	0.2133

TABLE 4.4: Spearman correlation coefficients for raters' *Familiarity* with native accents, for the "How sure" group.

Talker group	Raters' accent familiarity	Spearman coeff.
Native	British	-0.0811
Native	American	-0.0797
Native	Other native	-0.1382
Non-native	British	0.0356
Non-native	American	0.0340
Non-native	Other native	-0.0341

TABLE 4.5: Spearman correlation coefficients for raters' *Familiarity* with native accents, for the "How strong" group.

For hypothesis 4.4.2—familiarity with a non-native accent of English makes raters less able to distinguish it from native speech—we also tested the predictions that familiarity with the Czech accent in English will be inversely correlated with the accent *Score* given to Czech talkers of English, and that familiarity with the Slovak accent in English will be inversely correlated with accent *Score* given to Slovak talkers of English.

Spearman correlation coefficients were calculated for the combinations of talker native language (Czech and Slovak), accent familiarity, and the "How sure" and "How strong" groups. No correlations between Accent familiarity and accent *Score* were found.

4.4.3 Discussion

In this section, we tried to find out whether foreign accent *Scores* in our data set were correlated with the raters' familiarity with native and non-native accents in English. For *Familiarity* with non-native accents we did not find any correlation, which suggests that the knowledge of Czech- and Slovak-accented English neither made raters less sensitive to these accents nor gave them any significant advantage in spotting those accents. It may be, however, that any such (dis)advantage was covered up by other, more important factors, such as *Familiarity* with *native* accents. The data set, however, did not contain too many raters who specified low *Familiarity* with native accents and, therefore, this idea could not be investigated any further. An alternative explanation would be that *Familiarity* with non-native accents in fact means different things to different raters. For some it can mean that they are less sensitive to the foreign accent—this is especially conceivable in the case that they speak the accent themselves. For other raters it can mean that they are more aware of the foreign accent—especially if they speak a different accent themselves.

As for the results for native accents, it seemed surprising at first glance that the "How sure" and "How strong" groups should behave differently. Recall that in Section 4.1.3 we reported that the groups had overall very similar medians of *Score* values, and they behaved in a similar way also in terms of FAR in adverse listening conditions (Section 4.2.3), and in terms of correlation of *Articulation rate* and accent *Score* (Section 4.3.4). How was it possible, then, that the correlation between *Score* and *Accent familiarity* was stronger in the "How sure" group than in the "How strong" group? And why was the correlation between *Score* for native talkers and *accent familiarity* different for various native accents, but was more uniform when it came to *Score* for non-native talkers? First, there were just two native British English talkers and two native American English talkers, so even small differences between them could influence the correlation of *familiarity* with either British or American accents quite a lot.[21] Besides, the *familiarity* of British, American, and other native accents were correlated among each other (Spearman corr. coeff. > 0.35), therefore, perhaps we should have devised a combined metric for general native-accent *familiarity*.

More importantly, however, it seems that when it comes to *Accent familiarity*, the difference between the two accent rating tasks employed in our experiments suddenly becomes important. The answer to the "How sure" question captures the rater's ability to spot the accent—it is more rater-centric—whereas the "How strong" accent *Score*

21. For illustration, one of the British English talkers spoke a kind of Northern accent. She pronounced, for instance, the word "after" as /æft/ instead of the RP /ɑːft/ or GA /æːftɚ/, though this word appeared in one of the training stimuli that were not included in the main rating session.

expresses the talker's accentedness, which does not depend so much on the rater's *Accent familiarity*. If raters know the accent well, they can label it correctly as such even if it is weak—good *Familiarity* brings about high confidence *Scores* for all degrees of accents. This is where the positive correlation comes from. In contrast, if raters know the accent well, they can assign a low accent *Score* to the stimulus if the accent is weak, and a high accent *Score* if it is strong—good *Familiarity* brings about all sorts of *Scores* depending on the talkers. No correlation between *Accent familiarity* and accent *Score* should be expected in that case. In summary, it appears from our results that, unsurprisingly, knowing (any) native accents helps raters distinguish native from non-native talkers, but the new finding is that the rating task can be used to direct the raters' attention to more clearly differentiate between the two talker groups. Because of this, it seems more advantageous not to conflate these two kinds of rating into a single measure, as is sometimes the case in FAR experiments (see Section 4.1.3).

4.5 Language (mis)match between talker and rater

The question of native language (mis)match between talker and rater in FAR is connected with the question of *Accent familiarity*, but it is distinct. Even a rater who shares the native language with a non-native talker cannot automatically be expected to be familiar with the specific accent in the L2, as they may have learned the L2 in a different context. Even native raters and talkers can have a different experience with accents in their native language. Earlier studies have usually found, however, that native listeners can more reliably detect non-native talkers as such and tell them apart from native talkers. Non-native raters have sometimes been found to be more lenient with non-native talkers with the same L1, but not always (see Section 2.1). I wanted to see whether native raters in my data set benefited from their background knowledge, and whether non-native raters exhibited something like a "matched-interlanguage accent-rating benefit".

4.5.1 Method

From the raters in the FAR data set (see Section 6.2.6) we selected three groups based on the native language they specified in the language questionnaire. The "Native" group consisted of 53 native speakers of English, including the bilinguals (except for one bilingual in Slovak who was removed from the analysis because we only tested raters who fit into just one group). The "Czech" group consisted of 90 native speakers of Czech (including one bilingual in Polish; again, one Czech-Slovak bilingual was removed from the analysis). The "Slovak" group consisted of 70 native speakers of Slovak (including one bilingual in Hungarian). The remaining 109 raters were used as the basis for the "Other" group.

The Kolmogorov-Smirnov and Wilcoxon tests were used to compare the differences in probability distributions and medians when each group was compared to the rest

of the raters (including the other two groups). The "Other" group, therefore, was not as homogeneous as the other three groups (see Section 6.2.6 for more details).

When we compared all groups of raters together (English, Czech, Slovak, and Other), the difference between rater groups was measured with the Kruskal-Wallis test (`kruskal.test` in R) instead of the Wilcoxon and Kolmogorov-Smirnov tests because now we were comparing four groups instead of just two. As the Kruskal-Wallis test does not specify which and how many of the samples are different, the analysis was complemented by the Mann-Whitney U-test (`pairwise.wilcox.test` in R) to perform sample contrasts between individual subsets with corrections for multiple testing (Corder and Foreman 2009). For the four-sample tests, the boxplots were complemented by confidence intervals for median values in the form of notches in the boxes. If the notches in two boxes do not overlap, the differences can be considered statistically significant. We tested whether native and non-native raters fulfilled the following predictions in the two groups—"How strong" and "How sure"—separately:

Prediction 4.5.1: *Native raters will give a lower accent Score to native talkers than will non-native raters.*

Prediction 4.5.2: *Czech raters will give a lower accent Score to talkers who are native speakers of Czech than will raters who are not native speakers of Czech.*

Prediction 4.5.3: *Slovak raters will give a lower accent Score to talkers who are native speakers of Slovak than will raters who are not native speakers of Slovak.*

4.5.2 Results by language

Figure 13 shows that the Prediction 4.5.1 was confirmed. The native raters gave a lower accent Score to native talkers than did non-native raters, and the difference was statistically significant—Wilcoxon and KS tests for both the "How sure" and "How strong" questions returned p-value < 0.0001. The results were very similar in both. This corresponds to the general negligible difference between the groups (see Section 4.1).

Prediction 4.5.2 was also confirmed. Figure 14 shows that the results for the "How sure" and "How strong" groups corresponded to the general difference between the two: "How sure" ratings are on average higher than "How strong" ratings. The difference between Czech and non-Czech raters was significant in both groups—Wilcoxon and KS tests for both groups returned p-value < 0.0001.

Just like the previous two, Prediction 4.5.3 was also confirmed—as can be seen in Figure 15, Slovak listeners rated Slovak talkers as less accented than did non-Slovak raters. The difference was statistically significant in terms of both questions, "How sure" and "How strong"—Wilcoxon and KS tests for both groups returned p-value < 0.0001.

From these results it appears that the raters exhibited something similar to the so-called "matched-interlanguage intelligibility benefit" (see section 2.1), which refers

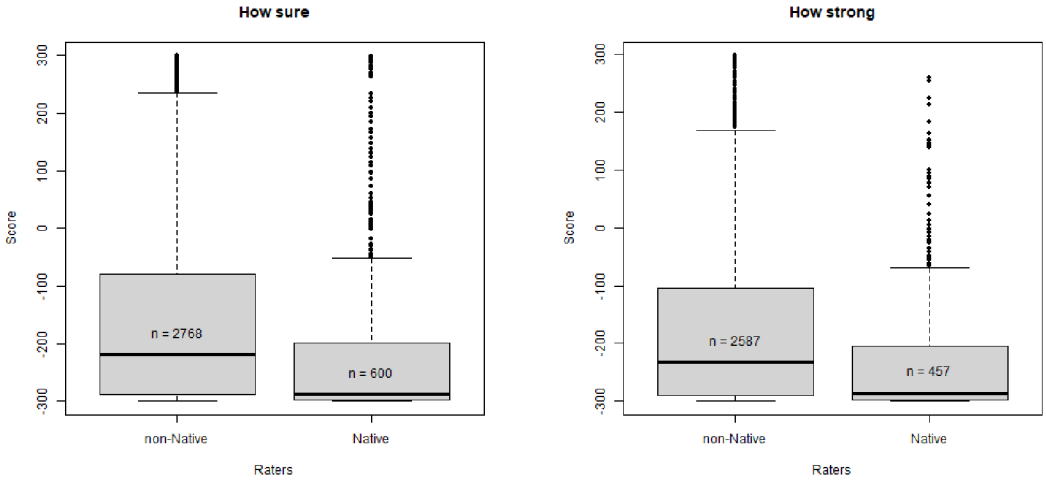


FIGURE 13: The difference between non-native (n = 269) and native (n = 53) raters when rating native talkers was statistically significant in both the "How sure" and "How strong" groups.

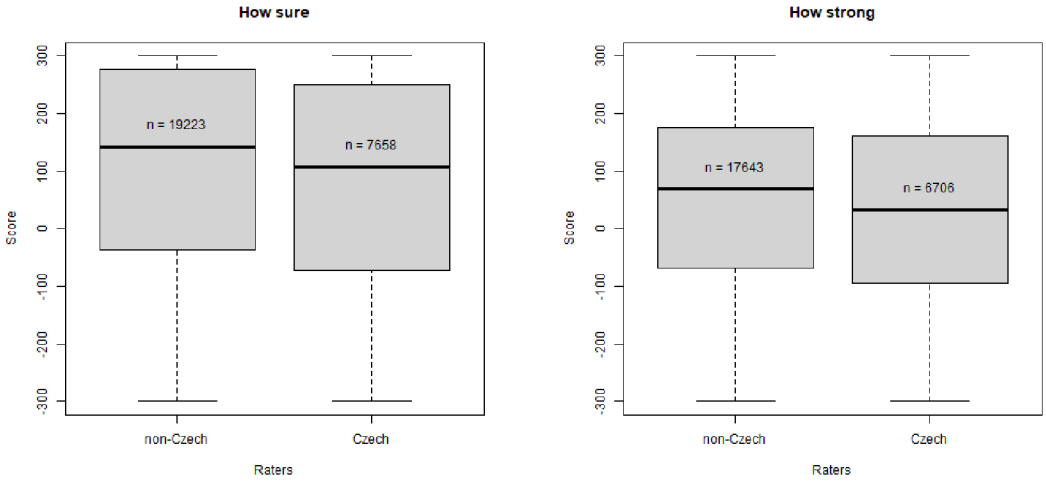


FIGURE 14: The difference between non-Czech (n = 232) and Czech (n = 90) raters when rating Czech talkers was statistically significant in both the "How sure" and "How strong" groups.

to the findings that non-native listeners can process non-native speech better if they share the native language. In our case, it seems that listeners rate talkers as less accented if they share the native language.

4.5.3 Results from all raters

Since Czech and Slovak are closely related languages, and the Slovak talkers in the data set were students of English in Olomouc, Czech Republic, and, furthermore, many Slovak raters in the data set also studied in the Czech Republic, we decided to have a more detailed look at how the "Other" groups[22] of listeners rated Czech,

22. As should be clear from the Method description, there was not just one "Other" group, because for each of the English, Czech and Slovak groups, the "Other" group also included the remaining two.

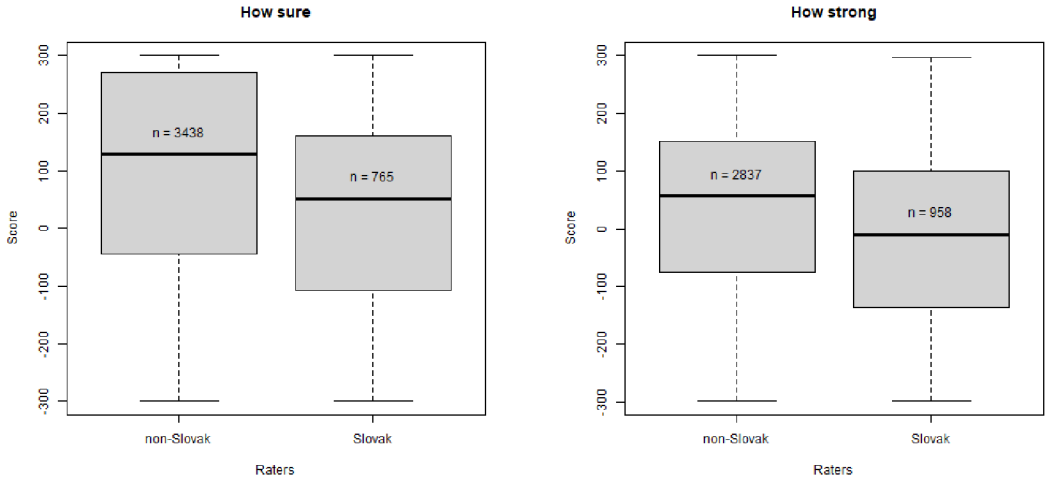
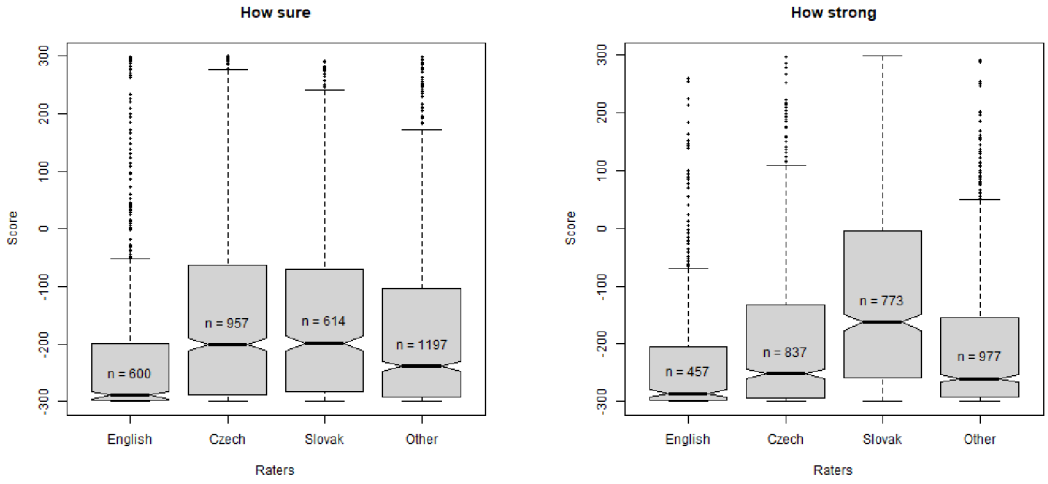


FIGURE 15: The difference between non-Slovak (n = 252) and Slovak (n = 70) raters when rating Slovak talkers was statistically significant in both the "How sure" and "How strong" groups.

Slovak and native talkers, in comparison to the respective native raters and we split the "Other" groups into four individual categories by native language.

The results of the Kruskal-Wallis test showed that the differences between the rater groups were statistically significant in all cases (p-value < 0.0001). The pairwise comparisons of the Mann-Whitney U-test also found significant differences between most of the groups at p-value < 0.0001, with a couple of exceptions which are noted in individual Figures 16–18.



(A) The Czech and Slovak ratings in the "How sure" group are not significantly different.

(B) In the "How strong" group the ratings are different except for the Czech and Other listeners.

FIGURE 16: Foreign accent ratings for four native English talkers rated by four groups by native language: English (n = 53), Czech (n = 90), Slovak (n = 70) and Other (n = 109). Non-overlapping notches in the boxes show significantly different means.

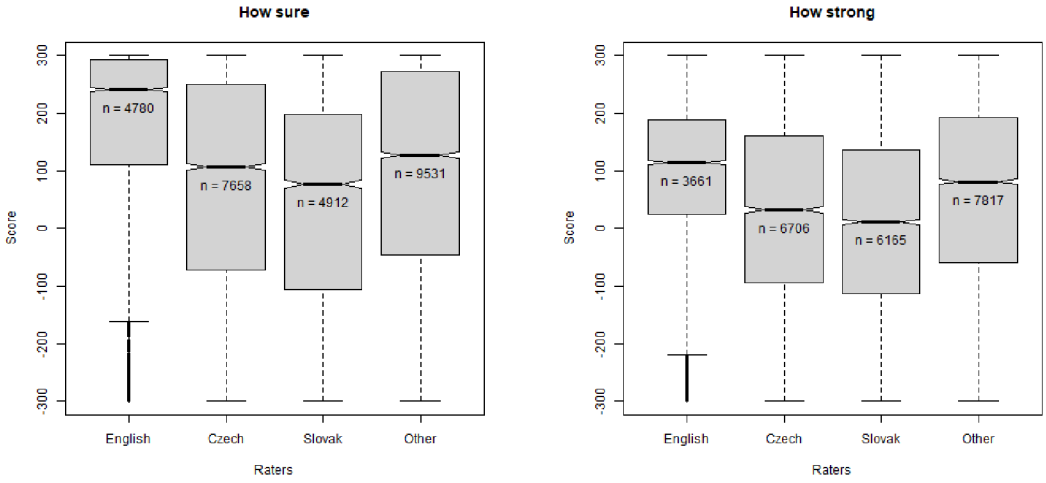
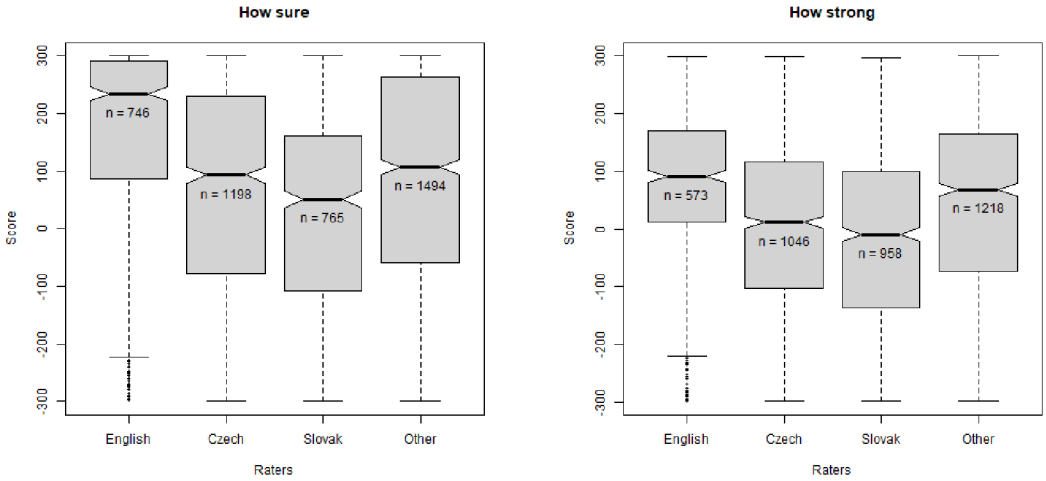


FIGURE 17: Foreign accent in 32 Czech talkers, rated by four groups by native language: English (n = 53), Czech (n = 90), Slovak (n = 70) and Other (n = 109). Non-overlapping notches in the boxes show significantly different means in all groups (p-value < 0.0001).



(A) Rating by all groups are significantly different from each oterh. The ratings by Czech and Other listeners for Slovak talkers on the question "How sure" are different with p-value = 0.011.

(B) In the case of the question "How strong" the difference between Czech and Slovak raters is different with p-value = 0.0023, and the difference between English and Other raters is different with p-value = 0.0022.

FIGURE 18: Foreign accent in five Slovak talkers, rated by four groups by native language: English (n = 53), Czech (n = 90), Slovak (n = 70) and Other (n = 109). Non-overlapping notches in the boxes show significantly different means.

4.5.4 Discussion

In this section, we addressed the question of how the native language (mis)match between talker and rater influenced foreign accent ratings. When looking at a binary distinction between raters with respect to the native language—native vs non-native, Czech vs non-Czech, and Slovak vs non-Slovak in Section 4.5.2—it seemed that listeners rated the talkers with whom they shared the native language as less accented

than did raters with a different native language. It appeared that the raters exhibited something similar to the so-called "matched-interlanguage intelligibility benefit" (see Section 2.1), which refers to the findings that non-native listeners can process non-native speech better if they share the native language and rate it as more intelligible. In our case, it seemed that listeners rated talkers as more native-like when they shared the native language.

However, when we took a more detailed look at the data, it turned out that this relationship did not hold true for all the language combinations. English native listeners gave the lowest accentedness scores to native talkers from all the raters and they also rated both non-native talker groups—Czech and Slovak—as more accented than any other rater group. Slovak raters, on the other hand, rated the native talkers as more accented than any other group of rates (with the exception of the Czech raters in the case of the "How sure" question), and they gave the lowest accentedness scores among all raters not only to Slovak talkers, but also to Czech talkers (Figure 17).

There results are in line with those of Wester and Mayo (2014) who found out that raters do not necessarily rate talkers with a matching native language as less accented than do raters with a different language background. The reason for this finding in our experiments may lie in the fact that the populations of Czech and Slovak raters and talkers in our data set were not really independent. As mentioned in the beginning of Section 4.5.3, many Slovak speakers are well familiar with Czech, and since the Slovak talkers, as well as a majority of the Slovak raters, were students of English in the Czech republic, they were also well familiar with the Czech accent in English. The results would have probably been different with more distinctly Slovak speaker and rater groups without a significant Czech experience.

5 Research questions related to biometric speech technologies

One of the points of the questions in Chapter 4 was to better understand foreign accent variation in the data set that will be used in evaluations of biometric speech technologies, which I will turn to in this chapter. We will look at technologies that were presented in Chapter 3, namely, Phonexia `LID` and `SID4`, and `lang-id` and `spkrec` by SpeechBrain. We will address the questions of how these technologies perform with non-native speaker data, how the performance of speech technologies is influenced by foreign accents, and how it is affected by the acoustic quality of the recordings.

5.1 Language identification, native language, and foreign accent

As an important part in many technological solutions, language identification technologies should perform well regardless of the native language of the speaker. In this section, we report the performance of selected language identification technologies—Phonexia `LID` and SpeechBrain `lang-id`—with the data set that was used in the foreign-accent rating experiments presented in Chapter 4, and with additional audio data from the same group of talkers. We considered the influence of recording length and recording quality on language ID accuracy, and looked at the effect of foreign accent in non-native English recordings. We analyzed the overall accuracy of the systems with the ACC metric (see Section 3.2), compared the score distributions, and analyzed correlations with other variables.

5.1.1 Method

For language identification experiments, 584 recordings in total were used, including all the English samples used in the accent rating experiments (336 samples from 42 talkers), and an additional 248 Czech samples from 31 speakers, which were used only in the speech technologies tests. The Czech samples were 4 phrases from the Czech version of "The North Wind and the Sun", each phrase in two versions, "Original" and "Phone" as described in Section 6.2.4.

The Phonexia `LID-L4` used in the tests was the same `3.42.1-lin64` version of `BSAPI` as `SID4-XL4`. The model has a recommended minimum speech length of 7 seconds, which means that, just like in the case of `SID4-XL4` (see Section 5.2.1), the model is not expected to work reliably with recordings that contain less speech, such as the recordings in the present set. The model distinguishes between 63 languages, including British English and U.S. English. In my experience, Czech speakers of English often combine features from various English accents, especially British and U.S. English. The scores for these two dialects were thus added up[23] and used as a single English score. Apart from that, the number of languages was limited to 45, to match the number of languages distinguished by the `lang-id` model.

By default, SpeechBrain's `lang-id` performs identification from 16 kHz audio files, so just like for `spkrec` a downsampled version of the data set was used (see Section 5.2.1 also for more information on sampling frequency and `BSAPI`). I also created a subclass of the `speechbrain.pretrained.interfaces.EncoderClassifier` that enabled me to do two more things: first, to save embeddings to disk in order to re-run the measurements without having to prepare embeddings anew every time; and second, to find out not only the best scoring language for each recording, but also the score for the actual language of the sample, e.g., if the language ID system returned "Swedish" as the identified language in an English recording, I also received the score for English.

5.1.2 Results: Native language and recording quality

The accuracy of speech technologies depends heavily on the similarity of training data to test data. Recording quality, such as "telephone speech" or "studio microphone", is one of the known crucial factors. Phonexia `LID-L4` is primarily trained on telephone data, whereas SpeechBrain `lang-id` is based on an open source data set recorded by a variety of devices. The two systems were therefore expected to perform differently with the two types of data in the present dataset. It is questionable how much non-native speech was included in the training data in either of the LID systems, but presumably, native data were in the majority and so we expected the following:

Prediction 5.1.1: *LID technologies will have a higher accuracy with native English recordings than with non-native English recordings.*

To see how the speech technologies performed with native and non-native stimuli, we started by selecting 336 English recordings (8 recordings per speaker) and pairing them with the best-matching language for each recording from the two language ID technologies. Based on the outcome, we calculated the Accuracy (ACC) as the ratio between correctly identified languages and the total number of recordings. Since the total number of native recordings was below 100, it did not make much sense

23. Technically, the scores were not simply added up, because log-likelihoods have to be treated in a special way, and so the "unified English score" was calculated according to this formula: $EN\_score = \log(\exp(EN\_UK\_score) + \exp(EN\_US\_score))$.

to compute percentages, hence the ACC measurements for these recordings have to be taken with reservation. With this in mind, we calculated the Risk Ratio (RR) to compare the Accuracy measurements for native and non-native speakers. In this context, RR is defined as the ratio between two probabilities of a correct classification. If the whole confidence interval of the RR value lies above 0, then the probability of a correct classification is higher in the first group. If, on the other hand, the whole confidence interval lies below 0, then the probability of a correct classification is higher in the second group. The results are shown in Table 5.1.

Talker group	`LID-L4` best match		`lang-id` best match		Total
	Correct (%)	Incorrect (%)	Correct (%)	Incorrect (%)	
Native	24 (75)	8 (25)	24 (75)	8 (25)	32
Non-native	175 (**57.6**)	129 (42.4)	226 (**74.3**)	78 (25.7)	304
Total	199 (**59.2**)	137 (40.8)	250 (**74.4**)	86 (25.6)	336
Risk Ratio	1.30 (95 % CI: 1.04–1.62)		1.01 (95 % CI: 0.81–1.26)		

TABLE 5.1: Accuracy of Phonexia `LID-L4` and SpeechBrain `lang-id` for native and non-native talkers is the percentage (values in parentheses) of samples correctly identified as English (values based on a reasonable amount of data are in bold). Risk Ratio values for native vs non-native speaker groups are complemented by their 95 % confidence intervals (CI).

Phonexia `LID-L4` performed better with the native recordings than with non-native recordings, with an RR of 1.30, which means that, it was 1.3 times more likely to classify the language correctly if the talker was a native speaker. In contrast, `lang-id` worked equally well for both kinds of recordings, meaning that, it worked better for the non-native samples than did `LID-L4`.

We further looked at how the LID technologies performed with a subset composed of Czech and English recordings of those 31 native Czech talkers who recorded samples in both languages during the recording sessions (see Section 6.1.1). In this case, the number of native-talker recordings was greater than in the case of native talkers of English, so the results had more weight. The analysis was performed separately for the "Original" and the "Phone" recordings to detect any influence of the "adverse listening conditions". We expected a number of outcomes based on the previous results:

Prediction 5.1.2: *For Phonexia* `LID-L4`*, the Accuracy will be higher with Czech recordings than with English recordings because Czech is the native language of the speakers, while English is their L2, and the system performed better with native English than with non-native English.*

Prediction 5.1.3: *For SpeechBrain* `lang-id-commonlanguage_ecapa`*, the Accuracy will be the same for Czech and English recordings.*

Prediction 5.1.4: *Accuracy of both systems will be higher with the "Original" recordings than with the "Phone" recordings because the "Originals" retain more spectral characteristics that can be used for creating a correct language model.*

The results are summarized in two tables: Table 5.2 shows the Accuracy overview of both LID systems for both types of recordings: "Original" and "Phone". Although it was not the main purpose of the test, we also compared the two systems with each other. Table 5.3 shows an overview of the Relative Risk values for four combinations of LID system and sample language in the two conditions: "Original" and "Phone".

Lang.	Condition	`LID-L4` best match		`lang-id` best match		Total
		Correct	Incorrect	Correct	Incorrect	
English	Original	71 (**57.3**)^{1,4}	53 (42.7)	105 (**84.7**)^{1}	19 (15.3)	124
Czech	Original	117 (**94.4**)^{2,4}	7 (5.6)	98 (**79.0**)^{2}	26 (21.0)	124
Total	Original	188 (**75.8**)^{6}	60 (24.2)	203 (**81.9**)^{7}	45 (18.1)	248
English	Phone	60 (**48.4**)^{5}	64 (51.6)	75 (60.5)	49 (39.5)	124
Czech	Phone	93 (**75.0**)^{3,5}	31 (25.0)	62 (**50.0**)^{3}	62 (50.0)	124
Total	Phone	153 (**61.7**)^{6}	95 (38.3)	137 (**55.2**)^{7}	111 (44.8)	248

Table 5.2: Accuracy of Phonexia `LID-L4` and SpeechBrain `lang-id` for English and Czech recordings of the same group of Czech native talkers (n = 31). Values in parentheses are percentages—Accuracy corresponds to the percentage of correct identifications. The upper indexes mark comparisons in the same column or row which are significantly different (see Table 5.3 and the main text for a commentary).

Comparison	Condition	Relative Risk	95% CI
`LID-L4` (En) vs `LID-L4` (Cz)	Original	**0.61**	0.52–0.71
`lang-id` (En) vs `lang-id` (Cz)	Original	1.07	0.95–1.21
`LID-L4` (En) vs `lang-id` (En)	Original	**0.68**	0.57–0.80
`LID-L4` (Cz) vs `lang-id` (Cz)	Original	**1.19**	1.08–1.32
`LID-L4` (En) vs `LID-L4` (Cz)	Phone	**0.65**	0.52–0.79
`lang-id` (En) vs `lang-id` (Cz)	Phone	1.21	0.96–1.52
`LID-L4` (En) vs `lang-id` (En)	Phone	0.80	0.64–1.01
`LID-L4` (Cz) vs `lang-id` (Cz)	Phone	**1.50**	1.22–1.84

Table 5.3: Relative Risk values for four combinations of LID systems (Phonexia `LID-L4` and SpeechBrain `lang-id`) and languages (English and Czech). An RR value above 0 means, the first system has a higher probability of returning the correct result for the language in parentheses, than the second system. Statistically significant values are in bold.

The results show that Phonexia `LID-L4` performed better with Czech data in comparison to its performance with non-native English data—in line with Prediction 5.1.2—and also in comparison with the performance of SpeechBrain `lang-id` with Czech recordings. Prediction 5.1.3 was also confirmed in that the performance of SpeechBrain `lang-id` was not significantly different in the two languages. SpeechBrain `lang-id` performed better with English data in comparison to Phonexia `LID-L4` Accuracy with English data.

The influence of recording quality could mainly be seen in that the superiority of SpeechBrain `lang-id` with English data diminished in the "Phone" condition, so that the two systems did not perform in a significantly different way with this language anymore. In contrast, the Accuracy of SpeechBrain `lang-id` with Czech data deteriorated in the "Phone" condition more than the Accuracy of Phonexia `LID-L4`

did, and so the upper hand of `LID-L4` with respect to Czech recordings became more prominent. However, when looking at the accuracy with all 248 recordings in both languages together, the better performance in one language and worse performance in the other evened out so that overall, the two systems did not perform in a significantly different way in either of the two conditions, "Original" and "Phone". Both systems, however, performed significantly worse in the "Phone" condition when compared to its performance with the "Original" data, RR of `LID-L4` was 1.23 (95 % CI: 1.09–1.39) and that of `lang-id` was 1.48 (95 % CI: 1.31–1.68).

5.1.3 Results: Amount of speech

As mentioned in the Method section above, Phonexia `LID-L4` is not expected to perform well with recordings that contain less than 7 s of speech. While the documentation of SpeechBrain `lang-id` does not explicitly specify any such limit, it is conceivable that it is also negatively affected by low amounts of speech in recordings. Since all of the 584 recordings that were subjected to the LID analyses were between 2.31 and 9.87 s long and contained between 2.05 and 6.29 s of speech (median: 3.39 s), according to the `VAD` submodule of `LID-L4`, we wanted to see how much the results depended on the amount of speech in the recordings. For this analysis, we calculated Spearman correlation coefficients for the random continuous variables *Amount of speech* and *LID scores* returned by the two technologies for the actual language in the individual recordings—either Czech or English.

All 336 English samples from the FAR experiments were used in this analysis, and an additional 248 Czech recordings from 31 Czech talkers (see Section 5.1.1). For both systems, a higher score denoted a higher probability that the language was spoken in the sample, so despite the difference in score range (see Sections 3.2.1 and 3.2.2), we expected the following for both systems:

Prediction 5.1.5: LID scores *will be positively correlated with the* Amount of speech *in individual samples.*

The results are presented in Table 5.4, which shows that both LID systems produced *LID scores* for all talker groups—Native English, Non-native English, and Czech—which were positively correlated with the *Amount of speech* in the recordings. In most cases, however, the correlation was not particularly strong. The strongest correlation (Spearman 0.53) was measured for `lang-id` in the case of native English recordings, however, there was not enough data to draw any strong conclusions. The results are also visualized in Figure 19, where the sparse scatter plots for native English show the limitations of the analysis. What can be seen quite clearly, though, is that both systems returned very high scores—and thus probably also the correct language identification—even for some extremely short recordings with *Amount of speech* 3 seconds or less.

Talker group	Spearman correlation coefficient	
	LID-L4	lang-id
Native English (n = 32)	0.1186	0.5352
Non-native English (n = 304)	0.2676	0.3157
Czech (n = 248)	0.0971	0.1501

TABLE 5.4: Correlation between the *Amount of speech* and Phonexia `LID-L4` and SpeechBrain `lang-id` *Scores* for four native and 38 non-native talkers of English, and for 31 Czech native talkers. Each talker was represented by 8 recordings, the n values in the first column denote the number of recordings in each group. All Czech talkers were also included in the English Non-native group, which also included the Slovak and Ukrainian/Russian speakers.

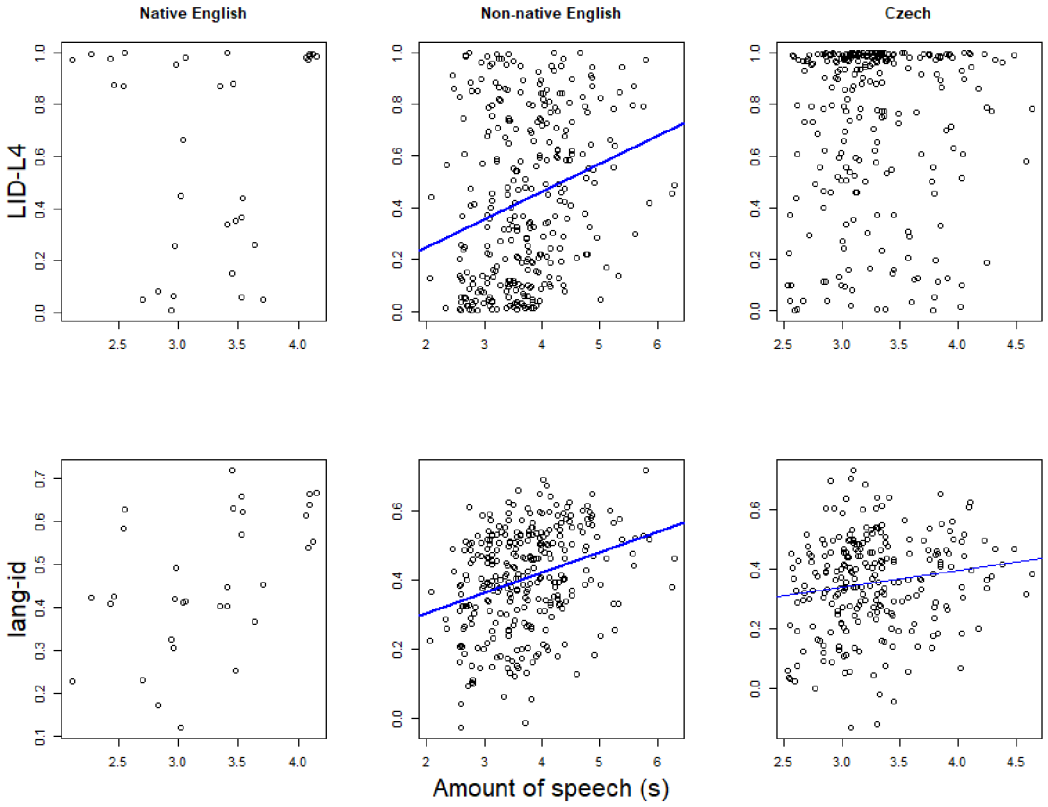


FIGURE 19: Significant correlations between *LID scores* and the *Amount of speech* for Phonexia `LID-L4` and SpeechBrain `lang-id` are marked by lines. The strongest correlation (Spearman 0.53) was measured for `lang-id` in the case of native English recordings, however, there was very little data to draw any strong conclusions, so we did not draw the correlation line in this case either.

5.1.4 Results: Foreign accent

An obvious extension to the question of whether the native language of the talker influences LID scores is the question about how LID scores correlate with the strength of foreign accent in non-native recordings. Since SpeechBrain `lang-id` performed the same with native and non-native English recordings (see Section 5.1.2), we did not expect that there would be any significant correlation between accent Scores and `lang-id` scores. In contrast, since Phonexia `LID-L4` had significantly worse results for the non-native samples—both when comparing non-native talkers of English with

native talkers, and when comparing native Czech recordings and non-native English recordings of the same talkers—we hypothesized that this was due to the phonological or phonetic deviations of the Czech-accented samples from native pronunciation forms, since native English likely predominated in the training data for the English language model of the `LID-L4` technology. We thus put together the *LID scores* and accent *Scores* for 336 samples from the FAR experiments, which were pronounced by four native and 38 non-native talkers, and we expected to find the following:

Prediction 5.1.6: *Phonexia* `LID-L4` *language ID scores for English will be negatively correlated with FAR Scores, whereas there will be no correlation between SpeechBrain* `lang-id` *and FAR Scores.*

The results are visualized in Figure 20. As predicted, we found a negative correlation between the `LID-L4` scores and the medians of accent Scores. The correlation was virtually the same in both the "How strong" and "How sure" groups, which is probably related to the fact that the results of the two rating groups were themselves strongly correlated (see Section 4.1.3. No correlation was found in the case of `lang-id`.

5.1.5 Discussion

In this section, we investigated three topics related to foreign accents and automatic speaker recognition technologies: the factors of native language, recording quality, and the amount of speech in recordings. It turned out that the performance of SpeechBrain `lang-id` did not differ significantly when comparing native or non-native data. Phonexia `LID-L4`, in contrast, was less accurate with non-native data, both when comparing native English with non-native English from two different talker groups, and when comparing native Czech and non-native English from the same group of talkers. We hypothesized that this could be due to the phonological or phonetic deviations of the Czech-accented samples from native pronunciation forms, which the system cannot cope with.

There were, however, two problems with this reasoning. First, the native English subset was rather limited (it only contained 32 recordings), and in general did not allow very strong conclusions. Brownlee (2020) illustrates how classification accuracy can fail as a metric in the case of highly imbalanced evaluation data sets. The other objection to our hypothesis is that we only tested non-native utterances in English. To better address the issue, we would at least have to also analyze non-native utterances in Czech to see if the lower accuracy was due to the non-nativeness of the talkers or rather due to a general deficiency of `LID-L4`'s English model. Assuming that `LID-L4` is trained to have a balanced performance for all languages, it seems more likely that the `LID-L4` model simply prefers native speech based on the training data. SpeechBrain `lang-id`'s English model, on the other hand, is trained from just one hour of English recordings (Sinisetty et al. 2021) (which may have been selected to represent a variety of accents), which would make it more robust when it comes to non-native data. Still another explanation might be that `LID-L4` is not well equipped

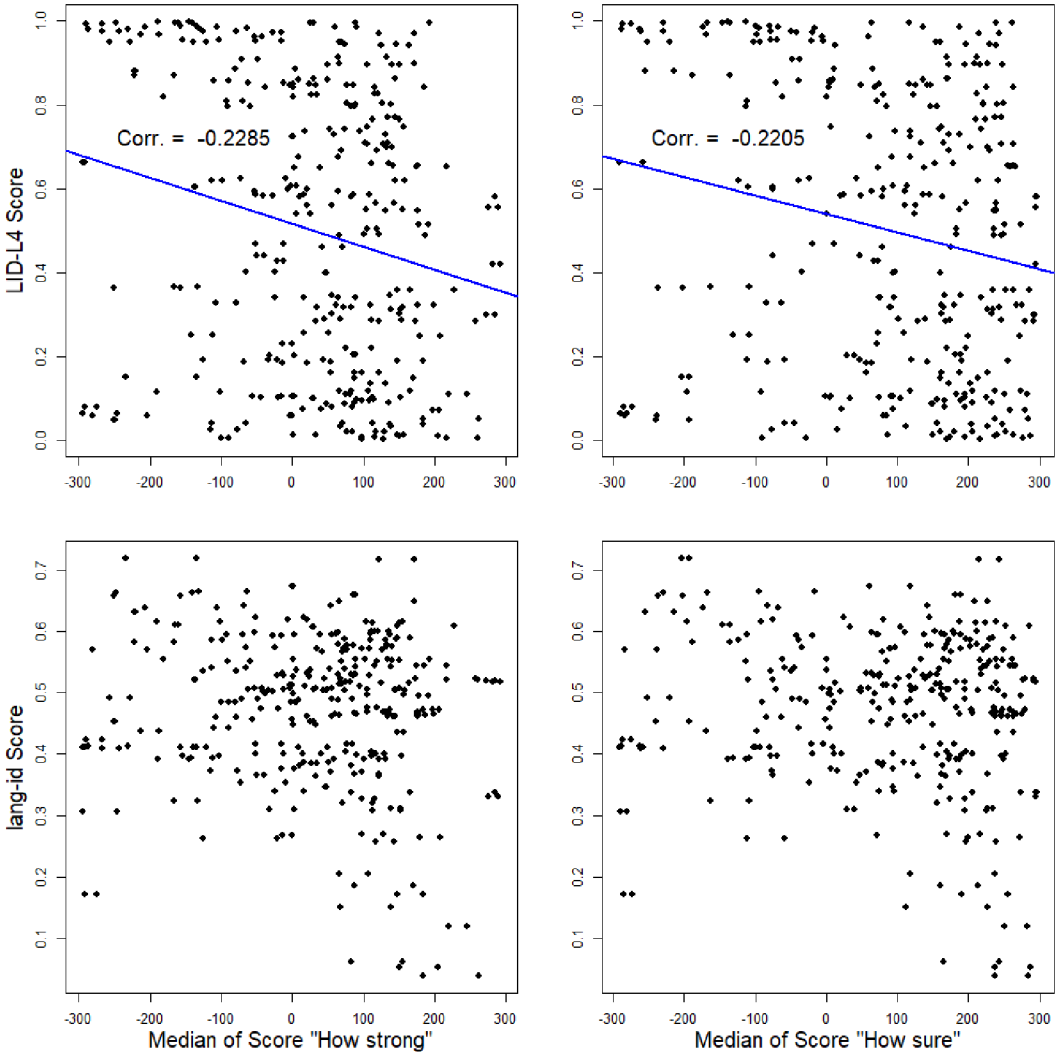


FIGURE 20: Correlations between *LID scores* and the medians of accent *Score* for 336 English stimuli pronounced by 8 native and 38 non-native talkers. Blue lines mark a significant negative correlation for the `LID-L4` scores. No correlation was detected in the case of SpeechBrain `lang-id`.

to cope with read speech, which is the type of speech found in both the data set used in the experiments reported in this thesis, and in the `lang-id` training data set.[24] However, the fact that the highest accuracy in our experiments was achieved by `LID-L4` with *read* Czech data indicates that the type of speech cannot be the only explanation, and may not at all be the right one.

Further, we examined the influence of recording quality on *LID accuracy*. We found that both `LID-L4` and `lang-id` performed significantly better with the "Original" recordings. This was not true only in the case of the `LID-L4`'s accuracy with non-native English data—which was not significantly lower in the "Phone" recordings—but only because its Accuracy with the "Original" non-native recordings was quite low to

24. To be able to investigate such questions was exactly the reason why part of the data collection procedure consisted of semi-spontaneous speech recordings (see Section 6.1.2). But the topic was not pursued further due to time constraints.

begin with. On the whole, Phonexia `LID-L4`'s suffered less from the detrimental effect of the "Phone" audio quality and, once again, this was likely because `LID-L4` is to a large extent trained on real telephone data with a sampling frequency of 8 kHz—the sampling frequency of the "Phone" recordings—and it automatically downsamples any input data to 8 kHz before further processing if it has a higher frequency—such as 44.1 kHz in the "Original" recordings. On the other hand, `lang-id` is mainly based on data recorded by a variety of microphones and, as mentioned before, its documentation strongly recommends to use 16 kHz. Again, this does not explain why `LID-L4` performed so well with Czech data in both conditions.

As for the *Amount of speech* as a factor in automatic LID identification, our analysis found weak positive correlations in all language groups and for both LID systems. The question of the amount of speech might seem a trivial one since more data can in general produce better results in biometric speech technologies. However, the correlations were not as strong as one might expect. A few things should be considered when looking for an explanation: First, the range of different amounts of speech was rather small: approximately 2–6.3 s. Most recordings would likely fall into the category of "very short" in case a data set with a larger range of amounts of speech was analyzed. Second, in the present data set, all the stimuli were the same set of phrases. While some talkers took more time to pronounce them, it did not really mean that they produced more speech with more syllables, and so more data for the system to analyze and make a better language model. Instead, longer recordings necessarily imply slower articulation rates, and we showed in Section 4.3.4 that slower articulation was linked to higher accentedness scores that, in turn, were linked to lower LID Accuracy.

5.2 Automatic speaker recognition and foreign accent

As outlined in Section 3.1.1, channel and language mismatches are some of the prominent challenges in automatic speaker recognition. Our FAR data set enabled us to test how two selected speaker recognition systems—Phonexia `SID4-XL4` and SpeechBrain `spkrec`—performed in such challenging conditions. In this section, we report a number of measurements that we performed in order to uncover the effects of language (mis)match and foreign accents on ASR accuracy.

5.2.1 Method

ASR scores were measured only for the 31 Czech native talkers[25] who recorded both English and Czech samples because, with these samples, the language mismatch could be systematically analyzed. Eight English phrases from each talker that were used in the accent ratings were complemented by eight corresponding phrases in Czech. In total, 496 recordings were used in the ASR tests.

25. As mentioned earlier, one of the male Czech talkers did not record data in Czech and was removed from the ASR tests.

As is recommended in ASR research and practice, only same-sex trials were analyzed (Doddington et al. 2000). Also, in the case of cross-channel trials (i.e., *Original* vs *Phone* recordings) the same recording was never compared to its alternative version, e.g., the original version of English Phrase 1 by talker 01 would never be compared with the phone version of the same recording, but was compared to other *Phone* recordings of the same talker.

Phonexia `SID4-XL4` can be provided as a command-line tool inside the so-called `BSAPI` (Brno Speech Application Programming Interface) software package. I used version `3.42.1-lin64` of `BSAPI` in the ASR tests. The model can create voiceprints (see Section 3.1.3) from very short audio files, even those that do not contain any speech; the module for voiceprint comparison (`vpcompare`), however, has a built-in limit of three seconds of speech (as determined by a voice activity detection submodule) as a protection from meaningless scores. This indicates that the technology is not expected to perform well with such short recordings. Nevertheless, since a number of the samples were shorter and did not fulfill this requirement, Phonexia provided me with a `vpcompare` application customized for testing purposes that did not have the limitation of amount of speech.[26]

The `SID4-XL4` model is trained on data with sampling frequency 8 kHz, as this is the sampling frequency in most telephone data that the technology is applied to and it performs automatic downsampling of input data before it is processed any further. So even if the original recording had a sampling frequency of 44.1 kHz, they were automatically downsampled to 8 kHz (this applies also to `LID-L4` and other components of `BSAPI`). The simulated telephone recordings still differed significantly because of the filtered-out frequencies and applied codes.

The documentation of `spkrec` cautions to make sure to use a sampling frequency of 16 kHz. The technology can process 44.1 kHz data, too, but it is unclear how that would influence the results, so all recordings were downsampled to 16 kHz for the purposes of `spkrec`.

5.2.2 Results: Language mismatch

In some use cases, an ASR technology may be required to identify the same speaker in two different recordings, even if he or she is speaking a different language in each. In such cases, technology users are frequently concerned over whether or not the ASR technology will perform reliably. As such, we wanted to find out what the overall accuracy of the ASR systems was when matching recordings in the same language compared to cross-language comparisons. Based on some previous experience with language mismatch in ASR, we put forward the following hypothesis and prediction:

26. An alternative solution would be to concatenate several recordings from individual talkers, e.g., join Phrase 1 with Phrase 2, and Phrase 3 with Phrase 4. This would, however, reduce the number of samples per talker and especially reduce the number of same-speaker trials, which would limit the information value of the evaluation.

Hypothesis 5.2.1: *The language of the samples being compared affects the accuracy of the ASR technologies because the system confuses the language similarity or mismatch for speaker similarity or mismatch.*

Prediction 5.2.1: *The ASR technologies will achieve a lower overall accuracy in terms of a higher Equal Error Rate for cross-language trials than for matched-language trials.*

The data set contained eight recordings in English and eight recordings in Czech from each of 31 Czech talkers; in both languages there were four "Original" recordings and four "Phone" recordings. We prepared ASR scores for all unique combinations among the recordings and calculated the Equal Error Rate to get a baseline measurement. We then divided the ASR scores into three groups and measured EER for each of them (see Section 3.1.2 for an explanation of EER). The first group contained only recordings in Czech, the second only recordings in English, and the third only cross-language trials. The results are shown in Figure 21 as two detection error tradeoff plots (DET). A DET graph plots against each other the false positive rates and false negative rates for all relevant thresholds and gives a complex overview of the accuracy. The axes of the plot are scaled in such a way that the lines in the graph are more or less straight, and the best system is the one whose line is the closest to the origin along its whole length. The plot is accompanied by the EER and EER threshold values. The best performance is typically the one with the lowest EER, and EER measurements are frequently used to optimize the parameters of an ASR system (e.g., to set the right threshold, select an appropriate calibration and normalization method) in order to achieve the best performance in a given use case.[27] What difference in EER is significant depends on the amount of evaluation data and on the practical implications in the use case. Considering the number of trials in our test, an EER difference larger than 1 % is almost certainly statistically significant (Richard Andrášik, personal communication).

Figure 21 shows that both Phonexia `SID4-XL4` and SpeechBrain `spkrec` suffered significantly from the language mismatch in the Czech-English trials when compared to the English-only trials. Furthermore, `spkrec` had a higher EER even for the Czech-only trials, which suggests that the language mismatch was not the only problem. As explained in Section 3.1.2, the EER measure is composed of both types of errors: false positives and false negatives. An ASR system can be more prone to one of the two—both disbalances increase the summarizing EER. If Hypothesis 5.2.1 is correct, we should find that the accuracy of ASR systems in the case of a language mismatch is held back by false negative decisions, and in the case of a language match by false positive decisions. The prediction can be stated as follows:

27. This is not necessarily the EER threshold but more often a threshold that brings about a certain false positive or false negative rate.

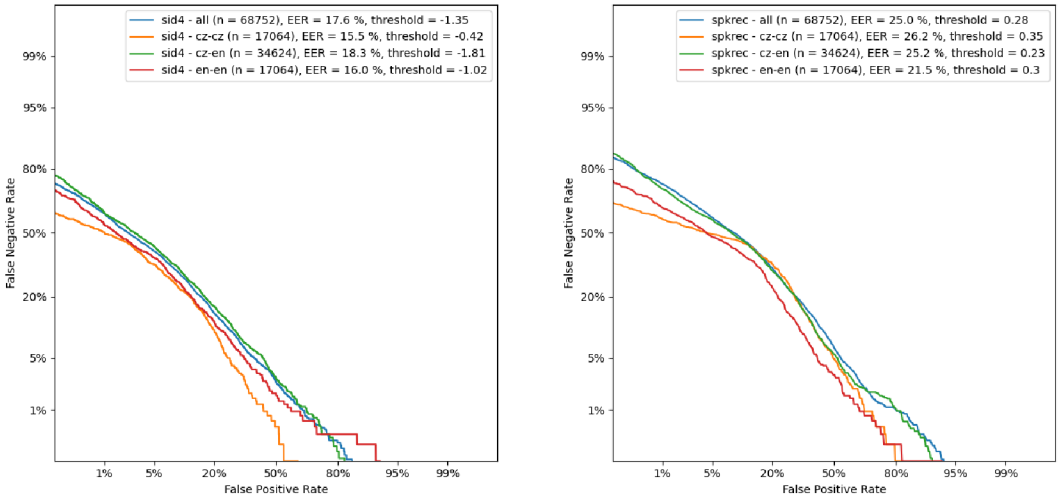


FIGURE 21: The error rates and EER values of Phonexia `SID4-XL4` and SpeechBrain `spkrec` in the language (mis)match conditions. The n values denote the number of trials in each condition: Czech-only, Czech-English, English-only, and "all" combinations taken together.

Prediction 5.2.2: *Matched-language trials (Czech-only and English-only) will get higher ASR scores than cross-language trials (Czech-English) when comparing not only different samples of one talker but also samples from different talkers.*

To test this prediction, we separated the trials into two groups—same-speaker trials, and different-speaker trials—because the groups should have different median values: positive and negative in the case of `SID4-XL4`, and greater than and less than 0.25 in the case of `spkrec` (see Sections 3.1.3 and 3.1.4). The results can be seen in Figure 22, which shows that Prediction 5.2.2 was confirmed: the cross-language trials received on average lower scores than the matched-language trials. The figure also visualizes a possible reason why `spkrec` had an especially high EER in the case of the Czech-only trials: the different-speaker trials received especially high scores, which lead to a high false positive rate.

5.2.3 Results: Foreign accent

It is likely that the effect of speaking one's native language in one sample and a foreign language in another will not be a uniform predictor of ASR accuracy with all speakers. We saw in Section 4.2.3 that some Czech talkers received FA ratings similar to those of native talkers. It is quite likely that a weak foreign accent in someone who can all but pass for a native speaker will have a different effect on ASR accuracy than a very strong accent. Does accent rating correlate with the accuracy of ASR technologies? When matching samples in different languages, are the ASR technologies less accurate with samples rated as more strongly foreign-accented than with samples with lower accentedness ratings? Or is it the other way round? We offer the following hypothesis and prediction:

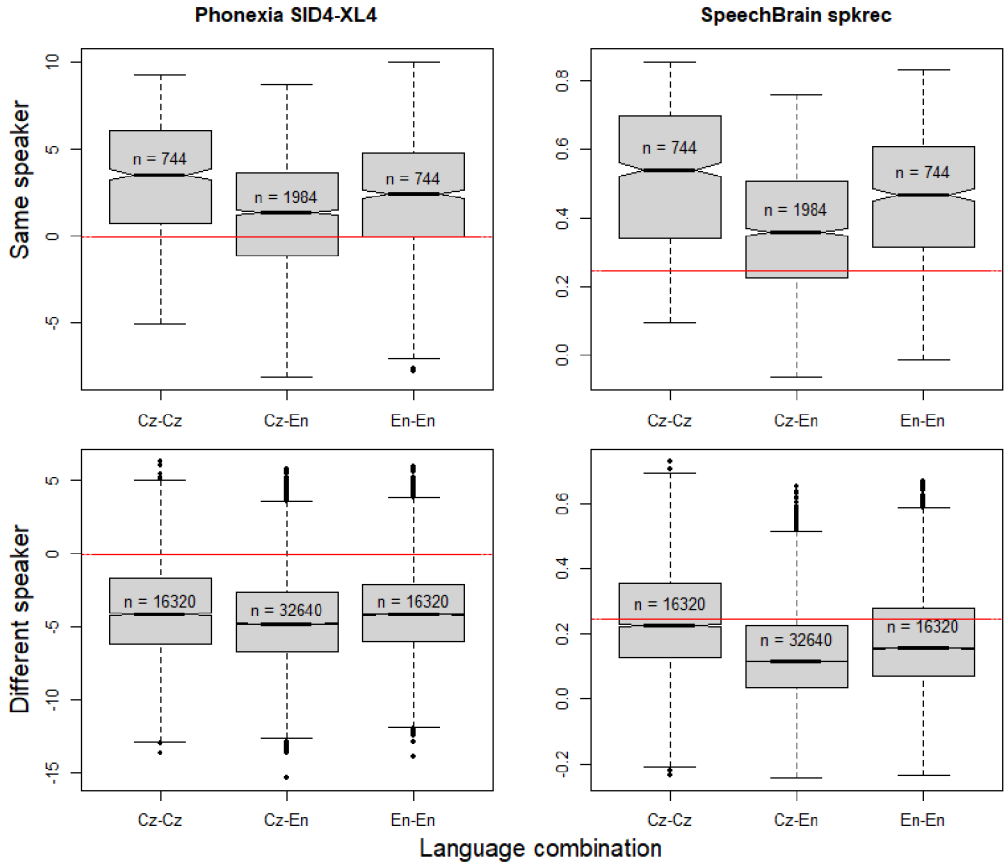


FIGURE 22: The median ASR scores correspond to the expected values with respect to the default decision points (shown as a red line at value 0 in the case of `SID4-XL4` and 0.25 in the case of `spkrec`, see the main text for more information). Notice, however, the high scores for the Czech-only different-speaker trials in the case of `spkrec`, which lead to the higher EER for this language combination. The groups within each subplot are significantly different from each other, according to the Kruskal-Wallis test (p-value < 0.0001), individual medians can be considered significantly different if the notches in the boxplots do not overlap.

Hypothesis 5.2.2: *A more native-like pronunciation of a speaker in a second language (represented by a low foreign-accent Score) makes it more difficult for the ASR technologies to recognize the speaker correctly in a cross-language context because the speaker's voice "timbre" in a well-mastered foreign language is different from that in his or her native language. In contrast, speakers retain more features from their native language in heavily foreign-accented speech, thus making it easier for the ASR technologies to correctly recognize the speaker.*

Prediction 5.2.3: *ASR scores for comparisons of Czech samples with Czech-accented English samples will be positively correlated with foreign-accent ratings received by the Czech-accented English samples. This will apply both to same-speaker and different-speaker trials.*

These groups were treated separately because, as was shown before, same-speaker trials in the data set had a positive median, whereas different-speaker trials had a negative median. The prediction was tested separately for the two accent scores, "How

sure" and "How strong", which focused either on the rater's confidence of perceiving a foreign accent or on the strength of the perceived accent. The results are presented in Table 5.5.

ASR model – Rating question	Same-speaker trials	Different-speaker trials
`SID4-XL4` – "How sure"	0.0964	-0.0423
`SID4-XL4` – "How strong"	0.1067	-0.0500
`spkrec` – "How sure"	0.2406	0.0628
`spkrec` – "How strong"	0.2635	0.0738

TABLE 5.5: Correlations (or the lack thereof) between *ASR scores* and accent *Scores* for non-native English recordings in the Czech-English cross-language condition (n = 34624). The two rating questions, and the same/different speaker groups were treated separately.

Based on the Spearman coefficients, there was apparently no correlation in the different-speaker trials. In the same-speaker trials, on the other hand, there seemed to be a weak positive correlation, especially in the case of `spkrec`. The Czech-English trials apparently received somewhat higher ASR scores if the English recording in the pair was rated as more accented.

5.2.4 Results: Channel mismatch

As we saw in Section 5.1.2, the recording quality (or channel, as it is typically called in the context of ASR) had a strong impact on the accuracy of LID technologies. It was only natural that the same factor should also play a role in the case of ASR technologies. The effect should not be as strong in the case of `SID4-XL4` as this technology is to a large extent based on the same training data and the same technology architecture as the `LID` technology. In the case of SpeechBrain `spkrec`, there is no such similarity of training data to `lang-id`, so the detrimental effect of the low quality in the "Phone" recordings is not so easily foreseeable. Nevertheless, we expected to find the following for both ASR systems:

Prediction 5.2.4: *The ASR technologies will achieve a lower overall accuracy in terms of a higher Equal Error Rate for cross-channel trials than for recording pairs in the same channel—"Original" or "Phone".*

In order to test the prediction, we measured ASR scores for all the unique pairs in a data set of 496 recordings from 31 Czech talkers (see Section 5.2.1). We divided the trials into two groups based on whether they contained only "Original" recordings, "Phone" recordings, or a combination, and calculated the EER values and DET plots. The results in Figure 23 show that the channel had a much stronger effect on ASR error rates than did language mismatch. Recall that in the case of the language mismatch, when both channel types were analyzed together, the best EER of `SID4-XL4` was 15.5 % in the case of the Czech-only trials, while `spkrec`'s best EER was 21.5 % in the case of the English-only trials. When both languages were taken together but the channels kept apart, the error rates were considerably lower. It seems that

for SID4-L4 Prediction 5.2.4 was marginally confirmed—the channel mismatch was connected with a lower overall accuracy in comparison to both subsets of recordings from the same channel. In the case of spkrec, a lower EER was only measured for the "Original" matched-channel trials; for the "Phone" recordings EER was even a little higher than for the mismatch condition.

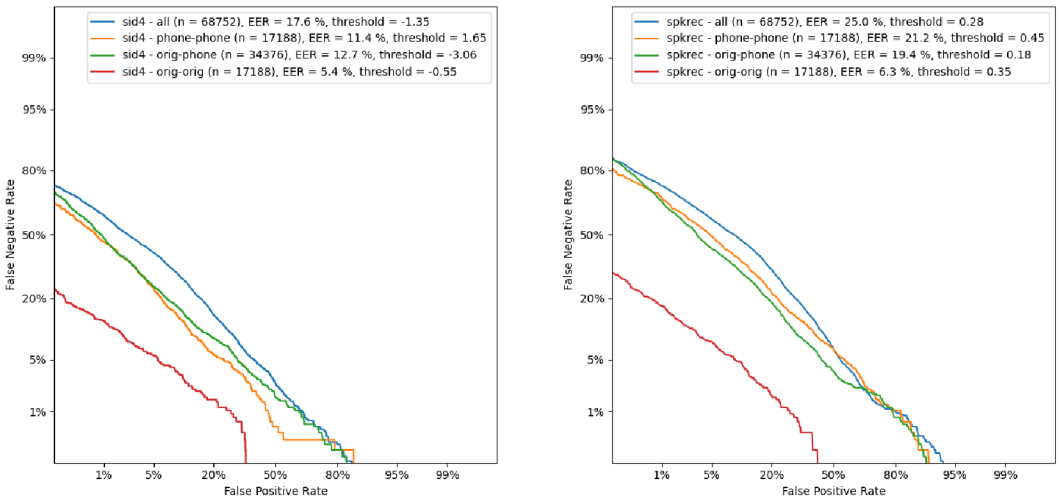


FIGURE 23: The error rates and EER values of Phonexia SID4-XL4 and SpeechBrain spkrec in the channel (mis)match conditions. The n values denote the number of trials in each condition: only "Phone", a combination of "Original" and "Phone", and only "Original". The values for "all" files taken together were of course identical to those in the case of the language (mis)match (Figure 21).

Just like in the case of the language (mis)match condition, we tested whether matched-channel trials received higher ASR scores than cross-channel trials. This was, like in the previous case, tested separately for the same- and different-speaker trials. Figure 24 shows that the exceptionally low "Original-Phone" same-speaker scores and the exceptionally hight "Phone-only" different-speaker scores were what mainly contributed to the high error rates in both ASR systems.

5.2.5 Discussion

In this section, we looked at three topics—language mismatch, foreign accent, and channel mismatch—in relation to automatic speaker recognition technologies. We found that the ASR error rates were dominated by the factor of channel. The "Phone" condition was especially challenging for SpeechBrain spkrec, likely because the technology was trained on 16 kHz data and expected input data in that format, while the "Phone" recordings were all downsampled to 8 kHz. Phonexia SID4-XL4 was not affected by the "Phone" condition as seriously since 8 kHz is the default sampling frequency it works with. In the case of SID4-XL4, the data seemed to confirm Prediction 5.2.4 in that the "Original-Phone" channel mismatch had a higher EER than both matched-channel conditions, however, the main difference occurred between

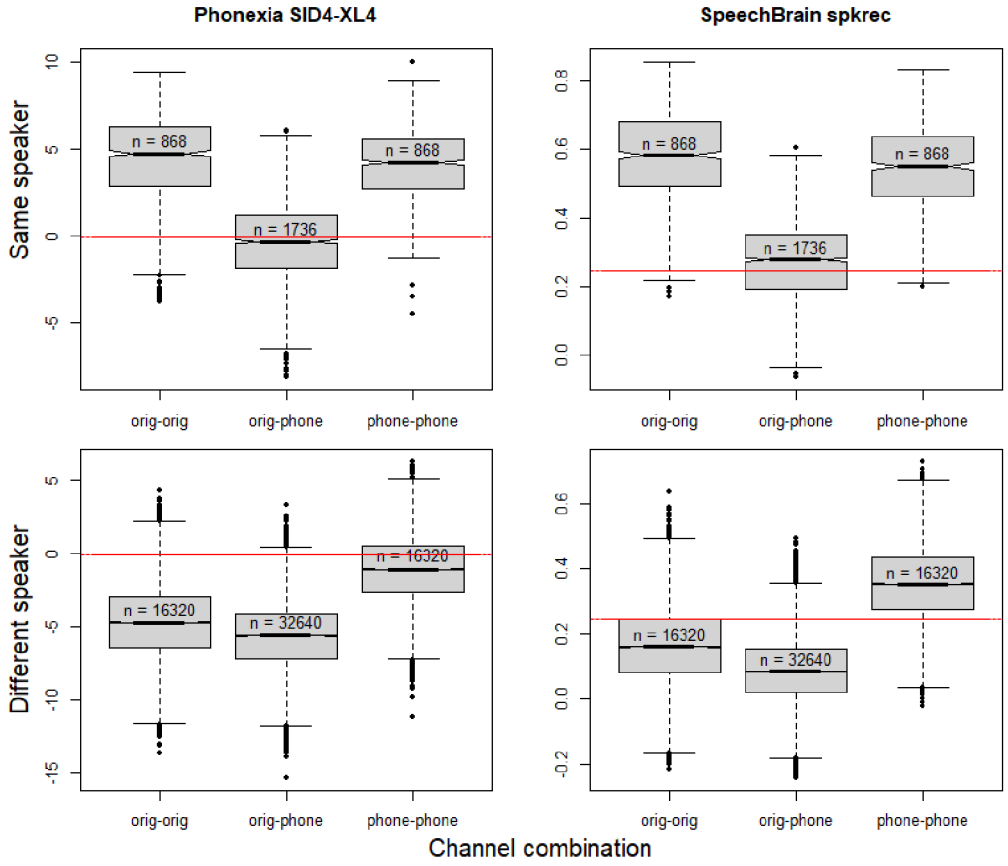


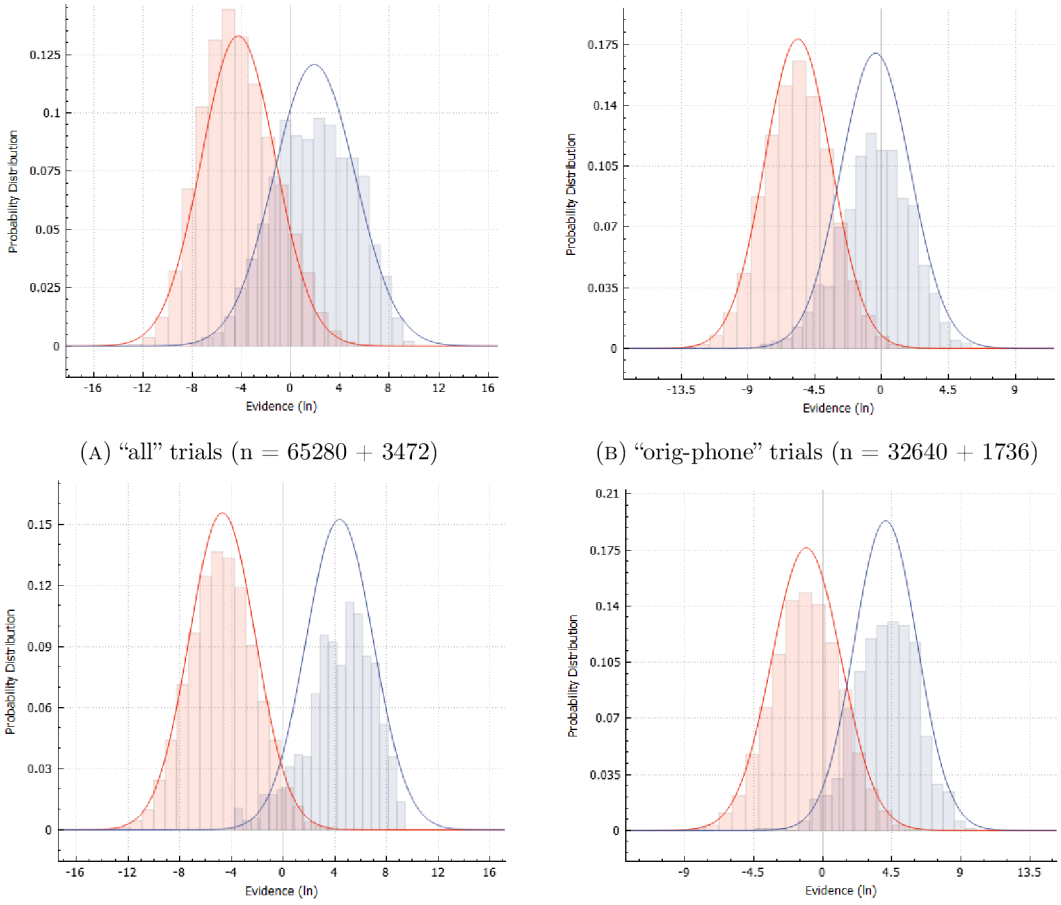
FIGURE 24: Channel (mis)match in the case of Czech-English trials. Not all median ASR scores corresponded to the expected values with respect to the default decision points (shown as a red line at value 0 in the case of `SID4-XL4` and 0.25 in the case of `spkrec`—see the explanation for Prediction 5.2.2 for more information): the scores for same-speaker "Original-Phone" trials are exceptionally low, while the scores for different-speaker "Phone-only" trials are exceptionally high. The groups within individual plots are significantly different from each other, according to the Kruskal-Wallis test (p-value < 0.0001).

the "Original" and "Phone" conditions per se. For `spkrec`, only the "Original-only" matched-channel condition produced a lower EER than the cross-channel conditions.

At first glance, it may seem a little surprising that the EER was higher when "all" recordings were taken together than when only recordings with a channel mismatch were compared (Figure 23). The reason is that error rates of individual subsets are not *averaged* when "all" recordings are considered together. Nor are they simply added, however. How the error rates are in fact combined together is visualized in Figure 25.

The four Probability Density Function (PDF) plots show the same-speaker and different-speaker scores as two histograms.[28] The area where the histograms overlap marks the errors made by the system. When a decision point is placed anywhere on the horizontal axis, all the same-speaker scores below the threshold are false negatives. Conversely, all the different-speaker scores above the threshold are false positives. If

28. Notice that the PDF plots show essentially the same information as the boxplots in Figure 24, only with the same- and different-speaker scores plotted on top of each other rather than in two separate plots.



(A) "all" trials (n = 65280 + 3472)

(B) "orig-phone" trials (n = 32640 + 1736)

(C) "orig-orig" trials(n = 16320 + 868)

(D) "phone-phone" trials (n = 16320 + 868)

FIGURE 25: The PDF plots show the distributions of different-speaker (red) and same-speaker (blue) scores (only `SID4-XL4` considered here). The n values show the number of different- and same-speaker trials in the test. The score distributions for "all" recordings combine score distributions from all three subsets, thus having the largest overlap of same- and different-speaker scores, and the highest error rate. The plots are screenshots from Phonexia's Voice Inspector forensic application, whose SID Evaluator can draw PDF plots from lists of scores.

the two score distributions overlap, there is no such threshold that would enable an error-free ASR system. By combining the score distributions from all subsets, the overlap increases and so does the Equal Error Rate.

This is likely the reason why the channel mismatch itself was not the biggest problem if it was the same in all trials. In the case of the "Original-Phone" condition (Figure 25.B), the mismatch clearly shifted all the same-speaker `SID4-XL4` scores down when compared to the "Original-only" trials (Figure 25.C), and increased the number of false negatives. In the "Phone-only" condition, in contrast, all different-speaker scores were shifted up (Figure 25.D), increasing the number of false positives. It was when these two contrasting score distributions came together that the highest error rates arose (Figure 25.A).

We saw that Prediction 5.2.1—about the role of language mismatch—presented in Section 5.2.2, was partially confirmed in that both ASR systems achieved the lowest

EER in the English-only trials. However, the error rates for the other matched-language condition—Czech-only—were not uniform across the ASR systems. The EER was lower than the EER for the Czech-English mismatch only in the case of `SID4-XL4`, while in the case of `spkrec` the EER for the Czech-only condition was even higher than the EER for the language mismatch. It is conceivable that the technology suffered from an effect similar to the *language-familiarity effect* observed in human listeners who have been found to more accurately recognize talkers in a language that they know. If Czech was missing from the Voxceleb datasets that were used to train `spkrec`, it might be more difficult for the system to recognize Czech acoustic features correctly as language-specific rather than speaker-specific.

Since the channel turned out to be such a strong factor in ASR accuracy, it likely overrode the effect of language. We thus looked at the language mismatch again, only this time with the two channels kept apart. Figure 26 shows that with the channel mismatch out of play, Prediction 5.2.1 was now confirmed for both ASR systems and for both channels: in both cross-language conditions—Czech-English, and "all" recordings taken together[29]—the ASR systems had higher EER values than in the Czech-only and English-only matched-language conditions.

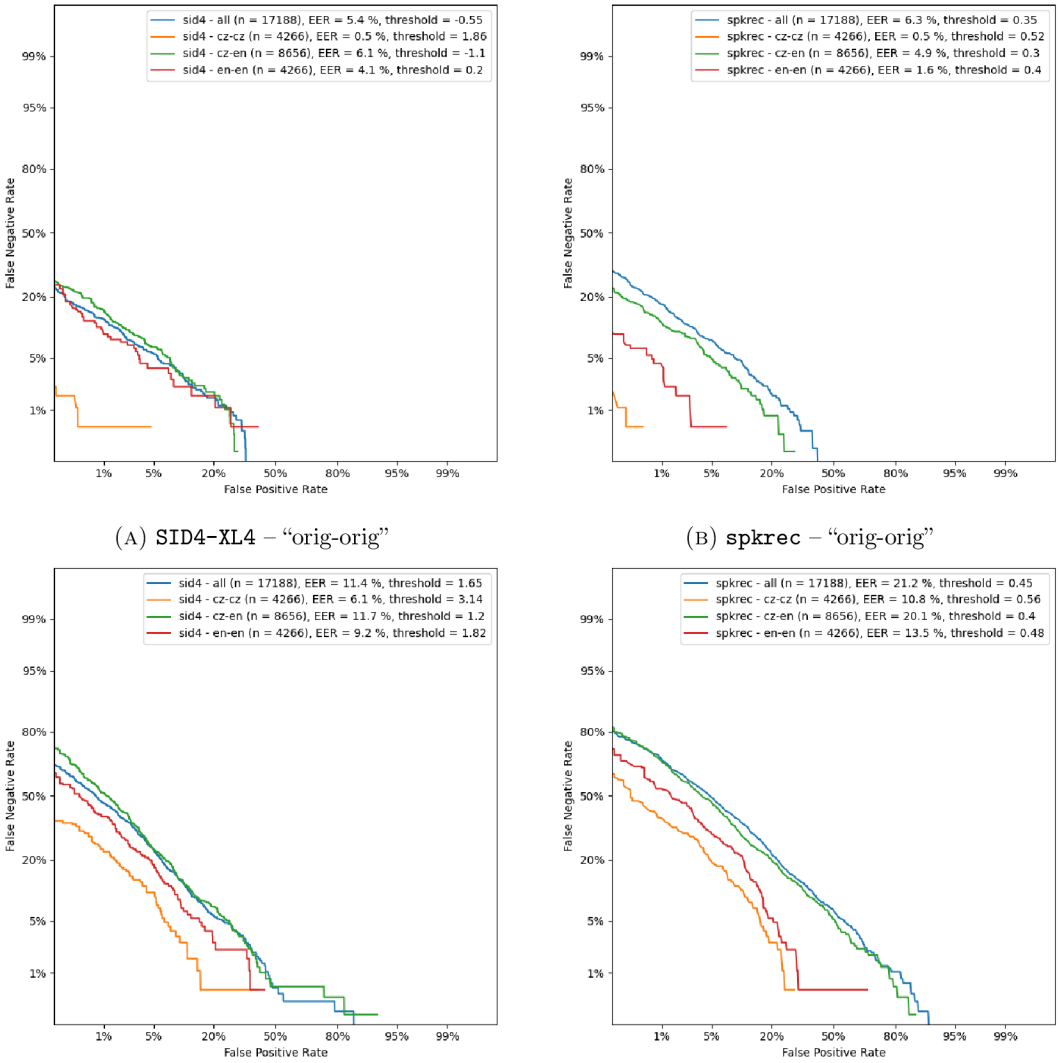
When both the channel mismatch and language mismatch were then controlled, both ASR systems achieved some very good results, especially in the Czech-only condition with an EER of 0.5 %. It thus turned out that `spkrec` did not have a particular problem with Czech data as we speculated earlier; the high EER of `spkrec` for the Czech-only trials shown in Figure 21 was more likely due to the interaction with the factor of channel. As was shown earlier in Figure 24, high Equal Error Rates were typically caused by unusually high different-speaker scores or unusually low same-speaker scores. To see what made the Czech-only EER so low in contrast to the other language combinations in the "Original" recordings (Figure 26.A-B), we broke the data up into same/different speaker scores in Figure 27.

The boxplots show that the above-average same-speaker scores in the Czech-only (and partially also the English-only) condition were what lowered the EER. However, the plots also remind us of the limitations of an analysis where individual groups of observations—divided according to several factors, such as channel, language, etc.—are treated separately: the number of observations can become very low, as in the case of the same-speaker trials.[30] Any attempts at an explanation of why the ASR systems were more accurate in the Czech-only trials—in comparison to the English-only trials—would have to remain in the domain of speculation, the more so because there was only one native language on trial.[31]

29. This was technically not a pure cross-language condition since the "all" condition also contained only a little more cross-language trials then matched-language trials.

30. The DET plots in Figure 26 also suggest a data paucity with the ragged curves, even though this is mainly caused by the fact that the same/different speaker score distributions hardly overlap in the Czech-only condition.

31. One such explanation would be the fact that Czech was the native language of the talkers and the ASR systems were presumably trained mainly on native-speaker data. Alternatively, it could have to do with the difference in articulation rate in the native Czech and non-native English samples.



(A) `SID4-XL4` – "orig-orig"

(B) `spkrec` – "orig-orig"

(C) `SID4-XL4` – "phone-phone"

(D) `spkrec` – "phone-phone"

FIGURE 26: The error rates and EER values in the cross-language conditions calculated separately for different combinations of channels seemed to confirm Prediction 5.2.1.

In Section 5.2.3 we saw that there was a weak positive correlation between FA ratings and ASR scores in the case of same-speaker trials. After finding the strong effect of channel on ASR scores, we revisited Prediction 5.2.3—ASR scores will be positively correlated with foreign accent ratings—only this time, we looked separately at "Original-only" and "Phone-only" trials.[32] We calculated Spearman coefficients for the correlation between the ASR scores and the FA ratings. The ASR scores were calculated for all Czech-English trials, and the median FA ratings for the English recordings within the pair were used. The results are presented in Table 5.6.

With the channel factored out, the positive correlations between the *ASR scores* and accent *Score* were noticeably stronger in the same-speaker trials when compared to the case when all channel combinations were considered together (cf. Table 5.5). In

32. Analyzing "Original-Phone" mismatch trials would be a little tricky. They could not be treated as a single group because it would have made a difference whether the English *or* the Czech recording would be the "Original" or "Phone" recording.

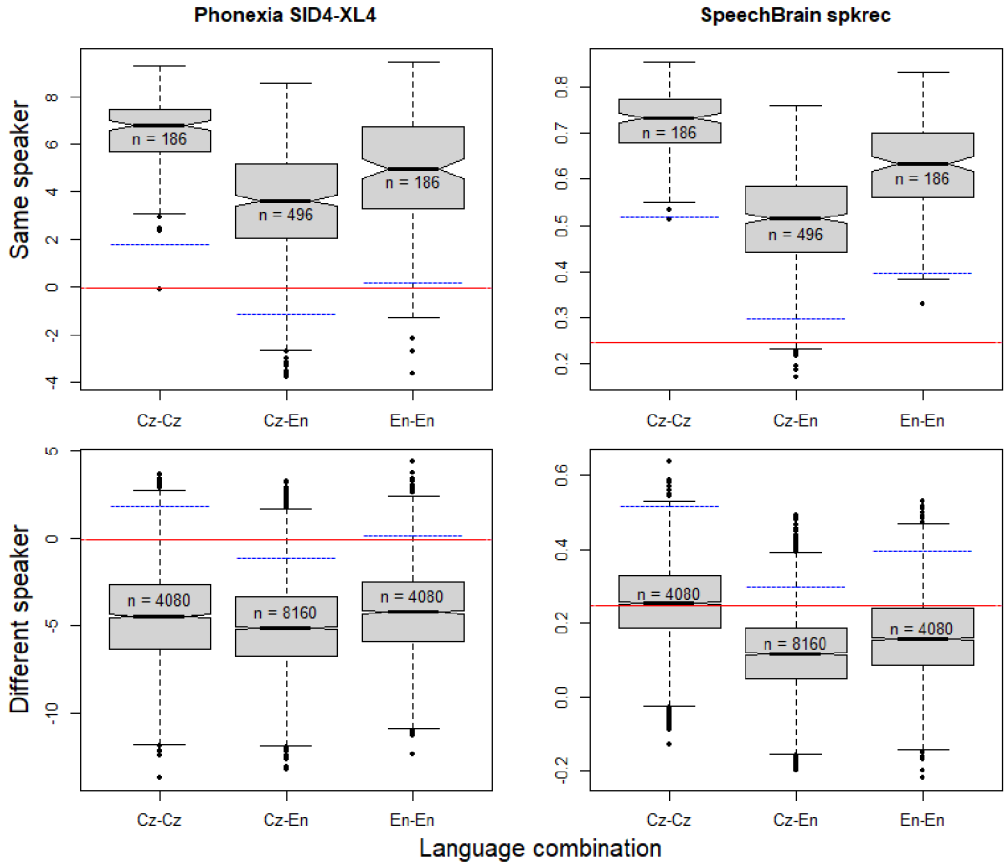


FIGURE 27: The boxplots break up the data in Figure 26.A-B into same/different speaker scores (only "Original" recordings are included). The red lines mark the ideal decision points (0, and 0.25 for `SID4-XL4` and `spkrec`, respectively), the blue lines show the EER thresholds for individual language combinations.

ASR model – Rating question	Same-speaker trials		Different-speaker trials	
	Original	Phone	Original	Phone
`SID4-XL4` – "How sure"	0.1777	0.2229	-0.0844	0.0234
`SID4-XL4` – "How strong"	0.2189	0.2001	-0.0740	-0.0071
`spkrec` – "How sure"	0.3189	0.3961	0.1299	0.0969
`spkrec` – "How strong"	0.3568	0.4263	0.1468	0.1203

TABLE 5.6: Correlations (or the lack thereof) between *ASR scores* and accent *Scores* for non-native English recordings in the Czech-English mismatch condition (n = 34624). This time divided into two subsets according to the channel.

the case of `spkrec`, the positive correlations were somewhat stronger even in the case of the different-speaker trials (but they were still weak). Prediction 5.2.3 was thus not fully confirmed in that a clear correlation between FAR and ASR scores was found only in the same-speaker trials, but not in the different-speaker trials. A possible explanation is that the Czech foreign accent is not something uniform or universal, which would make all Czech speakers of English sound more similar for the purposes of ASR technologies. It seems, though, that a strong accent can, to an extent, make

individual speakers sound similar in their native language and in an L2 so that ASR technologies can recognize them more accurately.

6 Data collection and experiments

This chapter is a detour into a description of how the data for this thesis were collected. If you are not interested in the details, you can skip to the concluding General discussion. The data used in this thesis can be divided into three main parts and were collected in several phases. First, there are the audio recordings and the metadata related to the talkers. The second phase of data collection consisted of an online accent rating experiment, which also included a collection of sociological metadata. The third data set used in this research is based on speech technology measurements performed on the recordings from the first phase of data collection.

6.1 Speech samples

6.1.1 Questionnaires

The speech data were collected using a H4n ZOOM recorder and a script[33] in Praat (Boersma and Weenink 2018), which guided the participants through several questionnaires related to language experience and through three tasks designed to elicit different types of speech data: reading and semi-spontaneous speech. The recording procedure was conducted by one of three assistants who explained the tasks to the participants, operated the recording device and started the Praat script.

Each talker was recorded in three different sessions, separated usually by at least one week in order to capture the natural variations of the human voice through time.[34] The first recording session started with a number of questionnaires that gathered the following information about the speaker: sex, age, highest level of education completed, field of study or profession, speech or hearing disorders, and details about their knowledge and use of languages.

Apart from English, participants were asked to record samples in their native language, and any other language they wished to.[35] Recording sessions always started

33. The script can be found at https://github.com/jakubbortlik/accent_rating.

34. The native English talkers were recorded under different conditions, because it was not possible to get them all in the recording booth in Olomouc. Also, their data was only intended to be part of the foreign-accent rating experiment, and not the follow-up experiments with speech technologies, and so they only did one recording instead of three, and not all of them recorded all three tasks which are described below, but they all recorded the reading task.

35. Originally, I was naïvely hoping to collect a larger multilingual data set but overestimated the linguistic competences of the students as well as my ability to figure out in time what to do with the data.

with the language for which the participant specified the highest proficiency. This was in most cases the native language, however, two participants, through the oversight of assistants, recorded German (Czech talker) and Czech (Slovak talker) instead. These talkers had to be removed from some of the experiments.

6.1.2 Speaking tasks

There was a short practice round in which participants learned the instructions for three tasks: reading, picture description, and spontaneous speech based on conversational topics. The participants were repeatedly asked to imagine they were talking to a friend on the phone, which was supposed to reduce the effect of the participants' unnatural isolation in the soundproof booth and to support the idea that the participant was taking part in a human interaction, not just delivering a monologue in front of a laptop. The Praat script created an annotation file which recorded the timing of the participants' interactions and the stimuli, which were shown on the screen in a randomized order. In all three tasks the participants had the possibility to pause the script in case they had any questions or problems.

The first task consisted of reading a modified version of Aesop's fable "The North Wind and the Sun". The participants had time to practice with the full text. Then each sentence was presented individually two times and the participants were instructed to read it aloud in a natural voice. There was a timeout of 15 seconds for each phrase, but since the phrases took just a few seconds to read, participants could practically continue at their own pace.

In the second task, the participants were instructed to describe in detail differences and similarities in five pairs of pictures. The pairs were completely different for each language, and there were modified versions of the pictures in each recording session. There was a time limit of 60 seconds for each pair of pictures and the participants could continue with the next one after 30 seconds.

The third task consisted of talking freely about five topics that were randomly selected from a list of about 200 conversational themes, such as hobbies, sports, countries, dugongs and other earthlings (see Monson 2005). There were 60 seconds for each topic, but the participants could go to the next one after 30 seconds and skip topics they did not like. The topics were unique across languages and sessions.

Only a small part of the whole data set was used in the foreign accent rating experiment (parts of the read phrases). I had plans for using the rest of the data for experiments with training new speech technology models but did not have the time to get it done in the end.

6.1.3 Recording devices and data format

After the practice round, the assistant checked the recording device and, in some recording sessions, also started a phone call that was recorded though the Phonexia Voice-Verify application. In this way, not only high quality studio data was captured

but also real telephone data transmitted over the network. The initial purpose of these telephone data was to test the FA rating in adverse listening conditions and the channel mismatch in speech technologies. However, due to technical problems, not enough telephone recordings were made. Moreover, it turned out that in some cases the ongoing telephone call created noise in the H4n recorder and rendered the "high-quality" recordings unusable, so the telephone recording was stopped.[36]

All recordings were made by the H4n ZOOM recorder (except for the two female native English talkers who were recorded by a laptop microphone). The original recordings were made with a 44.1 kHz sampling frequency and 24-bit precision in stereo format.

6.1.4 Talkers

The talkers were recruited primarily among students in the Department of English and American Studies at Palacký University. Some participants could earn extra credits in their course taught by Dr Šimáčková, but all participants were offered a chocolate bar and an analysis of their English samples by the Language ID technology developed by Phonexia. The native English talkers were either my friends, a colleague from work, or an English teacher known by acquaintance.

Thirty-two[37] talkers were native speakers of Czech, the second largest group were five Slovak native talkers, and one was bilingual in Ukrainian/Russian. Four control native talkers of English were selected to include one male and one female from both the U.K. and the U.S.[38] Table 6.1 summarizes the native language and sex characteristics of the talkers.

NL / sex	Czech	Slovak	Ukr/Rus	U.K. English	U.S. English
female	21	5	1	1	1
male	11	-	-	1	1

Table 6.1: Numbers of talkers by sex and native language (NL) represented in the data set. One of the 11 male Czech native talkers also indicated Slovak as another native language, but was treated as a member of the Czech group on the grounds that he spent most of his life in the Czech Republic and recorded material for the speech technologies experiments in Czech without any trace of a Slovak accent.

The native English talkers were included to define the "native" endpoint of the scale in the foreign accent rating experiment but they were not intended to be part of the technology analysis by speech technologies, mainly because of the difficulty in recruiting enough people. Since the native English talkers were not students anymore,

36. Alternatively, the phone data could have been used to create a "profile" for the equalizer in the VocalToolkit plug-in for Praat (Corretge 2019), which could be used to systematically process the original high-quality recordings and generate their telephone-like versions. Since there was not enough data for this, the default telephone profile from the VocalToolkit plug-in was used instead.

37. One of the 32 Czech native talkers also indicated Slovak as a second native language.

38. Originally, I intended to include more native talkers. I tried to contact international exchange students through several organizations at the university in Olomouc and in Brno but was unsuccessful.

they were all older (38–47 years) than the non-native talkers (19–26 years).[39] The ages of the talkers are summarized in this table:

sex \ age	NA	19	20	21	22	23	24	26	38	39	43	47
female	2	4	7	9	4	-	1	-	1	-	-	1
male	-	2	1	3	2	1	-	2	-	1	1	-

TABLE 6.2: Numbers of talkers by age and sex. Two female talkers did not specify their age.

6.2 Foreign accent rating experiment

The following sections describe the selection and preparation of samples for the rating experiment, the online accent rating procedure and the measurement of articulation rate.

6.2.1 Training stimuli

Two shorter phrases from "The North Wind and the Sun" were selected for training: Practice1: "But the harder he blew" and Practice2: "After that the sun began to shine". There were two versions of the training stimuli set. Each version contained only one phrase from each of the talkers: Practice1 was used from one half of the talkers and Practice2 two from the other half (42 phrases in total). The talkers in the groups were balanced with respect to sex and native language (and dialect in case of the native English talkers).

The purpose of the training stimuli was twofold: to make raters familiar with the rating procedure and to present them with the full range of accents they would encounter in the experiment. Some raters provided feedback that the training was too long and exhausted them too early, but it seemed necessary to present participants with a balanced representation of all talkers and not just a subset of the talkers or a random selection of the phrases.

6.2.2 Main rating stimuli

Four phrases from the story were selected for the main rating blocks:

1. The North Wind and the Sun were arguing which one of them was stronger.
2. Then a traveler came along wearing a heavy coat.
3. They agreed that whoever first got the traveler to take off his coat
4. And so the North Wind had to admit that the Sun was stronger.

39. Two Czech talkers did not specify their age in the questionnaire but they most probably fell within the range of 19 and 26.

The stimuli were selected similarly to (Volín and Skarnitzl 2010, 1013) based on the criteria that they contained a number of pronunciation features that are known to be problematic[40] for Czech speakers of English. See table 6.3 for more details.

The Czech talkers were expected to produce some combination of these foreign-accent features, many of which are also typical of Slovak English. One notable difference is that Slovak speakers of English can pronounce a final /v/ as [w] or [ʊ] as in "Do you love me?" [d uː j uː l a w m iː], which a Czech speaker of English would more likely pronounce as [d uː j uː l a f m iː]. The one Ukrainian/Russian native talker was expected to exhibit only a part of these features and to show a significantly different accent, especially with respect to stress patterns, vowel quality, and consonant palatalization.

Just like with the training stimuli, there were two versions of the rating stimuli set. Each version contained only two of the four phrases per talker. The phrases were paired in such a way that together they contained a similar amount of possible foreign accent triggers listed in table 6.3 above, so each talker was represented by either Phrase1+Phrase2 or Phrase3+Phrase4. There were 84 samples in each of the two main rating blocks.

The length of the samples was based on the considerations mentioned above in Section 2.6 and also based on some additional reasons. In general, individual short phrases were selected because, for most talkers, they formed compact intonational units that could be evaluated both on their segmental and suprasegmental characteristics. Since there was a rather high number of samples in the whole rating session (42 training + twice 84 main rating, totalling 210 samples), it was desirable to keep the sample length short, otherwise the experiment would be too demanding for the raters.[44] Also, individual phrases rather than concatenations of multiple phrases were used because, in this way, it was possible to collect ratings that reflected the variations of each talker. A larger number of shorter segments also enabled a more varied mixing up of phrases from individual talkers, which probably helped to counterbalance possible order effects.

Another thing that had to be considered with respect to stimulus length was that the samples were supposed to be fed into automatic speaker recognition later on. ASR technologies usually require a certain minimum amount of data to work reliably. The recommended minimum for Phonexia `SID4-XL4` is 3 seconds of speech, which is importantly not 3 seconds of *audio* but rather 3 seconds of what the submodule called

40. Only after the recording and rating experiment were finished did it occur to me that it may have been a mistake not to include non-problematic phrases in the rating, with few or no difficult sounds (if such phrases can be constructed) that might have more realistically represented the full range of each individual talker's accent variation.

44. I am afraid the experiment was unfeasible for many potential participants even in the final form, since it appears that, from the incomplete data saved by the PsyToolkit server, many people aborted the experiment prematurely. Unfortunately, no data were saved by the application in these cases. It might have been more "rater-friendly" to select only one, rather than two, of the four phrases from each talker, which would have halved the extent of the main rating phase. However, it would have also made the representation of talkers less balanced since the possible foreign accent features were not uniformly distributed among the phrases. It would also have possibly halved the number of ratings collected for each sample, although it is possible that more raters would have finished the experiment.

feature	(stereo)typical Czech accent in English	examples from "The North Wind and the Sun"
velar nasal /ŋ/	followed by a [k][41]	"wearing a" [w ɛː ɹ ɪ ŋ k ʔə]
dental fricatives	replaced by [s] or [t] in case of /θ/, and by [dz] or [d] in case of /ð/	"the, north, them"
word-initial vowels	preceded by a glottal stop and combined with devoicing of preceding voiced obstruent or, on the contrary, in the case of a *missing* glottal stop, combined with *voicing* of preceding *voiceless* obstruent (the former feature is more frequent in Bohemian Czech, the latter in Moravian Czech, and both can transfer into English Šimáčková, Podlipský, and Kolářová 2014)	"wind and" [w ɪ n t ʔ ɛ n t], "take off" [t eɪ g ɔ f]
labio-velar approximant /w/	replaced by [v] or [ʋ], correspondingly, /v/ can sometimes be replaced by [w] as a kind of "hypercorrection"	"Wind" [ʋ ɪ n t], "traveler" [t ɹ ɛ w ə l ə]
near-open front vowel /æ/	realized as [ɛ]	"traveler" /t ɹ ɛ v ə l ə/
open back vowels /ɒ/, /ɑ/	less open and back realizations as [ɔ] and [ä], respectively	"stronger" [s t ɹ ɔ ŋ g ə][42], arguing [ä ɹ g j u ɪ ŋ]
final unstressed /i/	realized as lax and short	"heavy" [ɦ ɛ v ɪ]
alveolar approximant /ɹ/	realized as a tap [ɾ] or trill [r]	"North", "arguing", "strong"
weak forms	full vowels, stressed syllables	"and", "of", "that", "were", "was",...
word accent	misplaced (typically to the first syllable, but occasionally from the first to another)	"along", "agreed", "whoever", "admit", "arguing" [ɛ ɹ ˈg j u ɪ ŋ k]
final obstruent voicing	voiced obstruent realized as voiceless, and vice versa (depending on the following sound), no pre-fortis clipping or pre-lenis lengthening	"wind" [w ɪ n t], "which one" [w ɪ dʒ w ʌ n], "first got" [f ɜː z d g ʌ t]
VOT	missing aspiration[43]	"coat" [k ɔʊ t], "take" [t eɪ k], "tighter" [t aɪ t ə]

TABLE 6.3: Overview of possible foreign accent features in Czech English, exemplified with data from "The North Wind and the Sun".

Voice Activity Detection is trained to recognize as speech. The four rated phrases contained 17, 14, 18, and 15 syllables respectively, and most talkers pronounced them in more than 3 seconds.

6.2.3 Sample preparation

The recorded audio data of each talker typically contained six versions of each phrase – two repeats in each of the three sessions. The original motivation to record each participant three times was to create non-contemporaneous data, which are best for evaluating ASR systems (see Morrison and Enzinger 2016). However, during sample preparation it turned out that many talkers did not manage to record all the phrases correctly in a single session, so samples from different sessions had to be used in the rating experiment.

The samples were extracted from the recordings semi-automatically with Praat scripts[45] and were manually checked that they did not include any major disfluencies beyond the general fluency of the talker, i.e., if the talker stuttered, made a filled pause, or made a noise the sample was rejected. By default, the second reading of the phrase in the second recording session was used because talkers were expected to be well familiar with the text by then. If there were some issues with it, other readings were considered.[46]

Only the stronger channel of the original stereo recording was used. The samples were modified so that they contained approximately 150ms of initial and final silence, and their amplitude was scaled to approximately equal loudness. The cleaned samples were converted from the original WAV format to FLAC, the Free Lossless Audio Codec, which reduced their size (so that they would all fit in the limited storage allowed by the PsyToolkit server) without compromising their quality.

6.2.4 Telephone call simulation

For the purposes of the adverse-listening-conditions experiment, alternative versions of all samples were created to simulate the characteristics of recordings transmitted over the telephone system. The simulation was based on the description of a similar procedure recommended as an example for the evaluation of forensic evidence (Enzinger, Morrison, and Ochoa 2016) and was done in the following steps:

1. Downsampling to 8 kHz

45. The Praat script Phonetic Corpus Builder https://github.com/jakubbortlik/pcb that I created for the purposes of a different project turned out to be quite useful once again.

46. A frequent problem, which was discovered only too late, was that talkers were too eager to push the "Next" button in the recording procedure, and the noise from the keyboard overlapped with their voice. Another problem was that in some sessions a telephone was used to capture the voice together with the H4n recorder. When the two devices were too close together, the H4n audio was corrupted by stationary noise. One session of one talker was recorded in a different room and there was too much echo so this session had to be discarded.

2. Band-pass filtering with the VocalToolkit plugin in Praat (Corretge 2019), which filtered the lower and upper frequencies that are not present in telephone audio data (retained were the frequencies in the range 300–3400 Hz)

3. Encoding into a-law (a compression format frequently used in telephony) and back to PCM, or pulse-code modulation, the standard format for audio signals (Smith 2007)

4. Encoding and decoding using the G.723.1 coding scheme (International Telecommunications Union 2006)

6.2.5 PsyToolkit experiment

In this section I describe how I used the free PsyToolkit platform (Stoet 2017) to run an online accent rating experiment. PsyToolkit made it possible for me to collect a relatively large data set from raters around the globe, but some of its features did not allow me to have full control over the data collection procedure. The toolkit offers two relatively simple scripting languages for creating online[47] questionnaires and reaction-time experiments and offers free hosting of experiments on a private server.

I wrote a PsyToolkit script[48] that ran an experiment with the following characteristics. The rating session started with a practice round that contained one shorter recording from each of the 42 talkers and was accompanied by more explicit instructions. The first screen explained the task and stated the rating question (Figure 28).

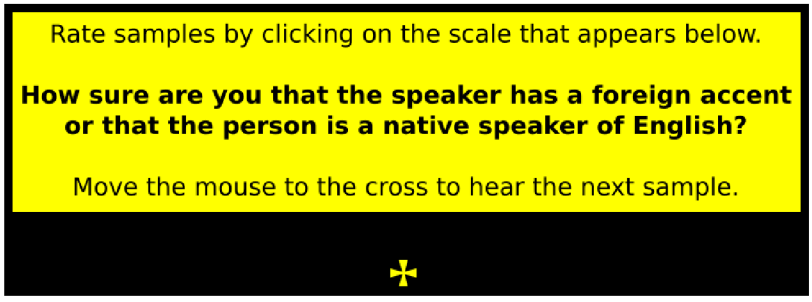


FIGURE 28: One version of the PsyToolkit rating experiment's instructions. Screenshot from a training round with more explicit guidance. The main rating round only included the text in bold. The sound was played automatically after reaching the cross with the mouse pointer.

When raters moved their mouse to the "fixpoint", a randomly selected sample was played automatically, and on the screen appeared the rating scale with labels at both ends and a replay button. The fixpoint was placed so that the center of the scale appeared right beneath the mouse pointer. The sound could be replayed once; after that the button turned gray and became inactive. Raters could click the scale as many times as they wished, updating the pointer each time. After clicking the scale for the first time, the "next" button appeared below the scale (Figure 29).

47. There is also an offline version for GNU/Linux (Stoet 2010).
48. The script and additional files are available at https://github.com/jakubbortlik/accent_rating

There was virtually no time limit[49] to press the next button or click the scale so that raters could proceed at their own pace. Reaction times were measured between playing the sound for the first time and clicking the scale and then the "next" button in order to discard any ratings that would be too separated from actually hearing the sample. Unfortunately, clicking the "replay" button was not timed so that during the processing of the data there remained some uncertainty as to whether the ratings of replayed samples were valid. However, raters replayed the samples relatively infrequently (about 14 % of the time), and in most cases clicked the "next" button soon enough even with respect to the first playing of the sample so that the sound probably stayed in their working memory (see Section 6.2.7).

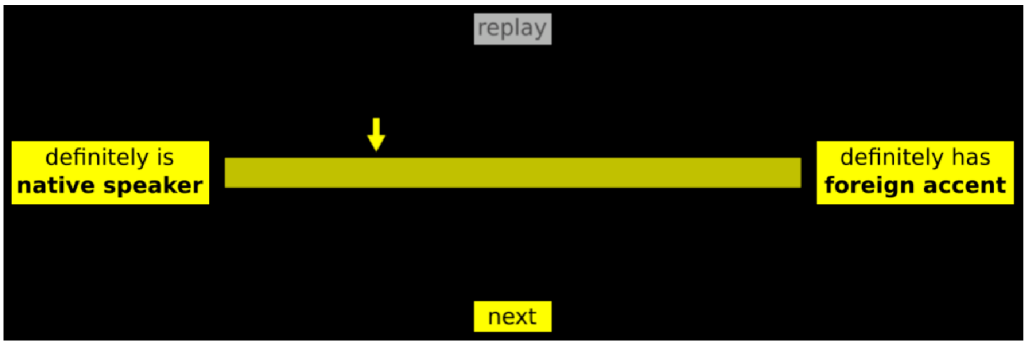


FIGURE 29: One version of the PsyToolkit rating experiment's scale after replaying the sound (gray button) and clicking on the scale (pointer marking the place where scale was clicked).

The practice round was followed by two main rating blocks separated by pauses when the raters could rest. The two main blocks contained identical stimuli to allow validation of intra-rater reliability; the order of the samples, however, was randomized each time.

In all versions of the experiment, the scale consisted of 601 vertical one-pixel wide lines, which were placed right next to each other so that they appeared like a continuous horizontal yellow bar. The scale was centered in the middle of the screen and reached 300 pixels to the right and 300 pixels to the left of the central line, which was not marked. Likewise, there were no marks along the scale. The central line corresponded to score 0—clicking on the "native speaker" part of the scale resulted in negative values, and the "foreign accent" part yielded positive values. The orientation of the scale ("native > non-native" or "non-native > native") was chosen randomly for each rater and did not change during the experiment.

There were two versions of the experiment to address the research question of rating task formulation. The two versions were assigned to raters randomly:

Rating task version 1: rater confidence

Raters had to answer the question "How sure are you that the speaker has a foreign accent or that the person is a native speaker of English?" The end points of the scale were labeled "definitely has foreign accent" and "definitely is native speaker".

49. There was, in fact, a timeout of one hour, which was only exceeded a handful of times by some participants who apparently abandoned the rating for some time and finished it later.

Rating task version 2: accent strength
In this version, raters had to answer the question "How strong is the foreign accent of the speaker or how much does the person sound like a native speaker of English?" The end points were labelled as "very strong foreign accent" and "sounds like a native speaker".

In general, I tried to minimize the amount of instructions the participants had to read and remember because the experiment was demanding enough as it was. I decided, therefore, not to inform raters in my experiment about the language background of the talkers because there were native English talkers as controls and not a very balanced combination of non-native talkers Czech, Slovak, and one Ukrainian-Russian bilingual.

6.2.6 Raters

The rating session was followed by a questionnaire that asked raters about their age, sex, native language, parents' native languages, knowledge of other languages (primarily English and the native languages of the talkers), familiarity with native and non-native English accents, the device used for the rating experiment, and how they learned about the study. Raters could also give feedback about the experiment.

Raters were recruited among friends, acquaintances, colleagues, and via email at universities in Czechia, Slovakia, Poland, Austria, Germany, the United Kingdom, and the United States. The mailing campaign mainly targeted English departments and humanities, but also included science departments (maths, physics) to "balance" familiarity with linguistics and with English accents. Advertisements were also published at LinguistList, and the PsyToolkit user forum.

Participants could receive a simple analysis of their rating results (a graph of average ratings per native language group among the talkers). Some participants—mostly Czech and Slovak students at the Technical University in Liberec—were promised extra points in their course if they participated. This turned out to be the most reliable way of recruitment; the mailing campaign, on the contrary, proved to be rather ineffective: I sent out nearly 190 emails to universities in English speaking countries and only about 50 native English raters decided to take part in the experiment (many of which learned about the study by different means than from their university teachers: friend's recommendations, social networks, etc.).

Most raters were monolinguals—only a minority indicated that they were bilingual. The native languages of the raters are summarized in Table 6.4.

Most raters were university students—especially the Czechs and Slovaks who were mainly recruited at the Technical University in Liberec—so the majority were in their early twenties. Since the experiment was advertised in many places, and the only limitation for participation was a minimum age of 18, the complete range of ages was quite wide, as can be seen in Table 6.5, together with the proportions of men and women. For likely similar reasons, the majority of participants were women because

Language	Number of raters
Czech	89 + 1 Slovak, 1 Polish
Slovak	69 + 1 Hungarian, 1 British English
German	52 + 1 Norwegian, 1 Italian
American English	32 + 1 Polish, 1 Marwari
Polish	20
British English	14 + 1 South African English
Russian	6
Hungarian	4
Arabic	3
Canadian English	2
Australian English	2
Ukrainian	2
Other	20
Total	324

TABLE 6.4: Number of speakers of different native languages among the raters. Numbers and language names after the plus sign indicate bilinguals. These are only included in the first group they belong to. The "other" group contained one or two native speakers of Spanish, Finnish, Dutch, Croatian, Vietnamese, Urdu, Turkish, Japanese, Italian, Indonesian, Greek, French, Macedonian, Chinese, Bulgarian, and Bosnian.

there are more female university students than male, at least in the Czech Republic (Český statistický úřad 2020).

sex \ age	18-19	20-29	30-39	40-49	50-59	60-69	79	total
Female	32	124	33	20	9	5	-	223
Male	12	47	23	7	3	3	1	96
NA	1	4	-	-	-	-	-	5

TABLE 6.5: Numbers of talkers by sex and age. Five participants did not specify their sex or used a non-binary option. In total there were 324 raters.

The majority of raters indicated that they used a computer with headphones (183) and a smaller part used a computer with speakers (141). Most raters were able to finish the experiment in under an hour.

6.2.7 Processing of rating data

Two raters were removed from the data set before their input was used in any analyses. One of the two reported a hearing loss of 70–80 %. The other rater was much more inconsistent in his ratings than any other rater. Rater consistency was verified by calculating the median of differences in *Scores* given by the rater to the same recording in the two rating blocks. Figure 30 shows an overview of the differences in ratings.

The reaction time (RT) between playing the stimulus sound and clicking on the rating scale was used as a rough indicator that the rater still had the sound in their auditory memory at the time of rating (see Zimmermann, Moscovitch, and Alaina 2016). In cases when the replay button was not *clicked*, ratings with an RT of 10s

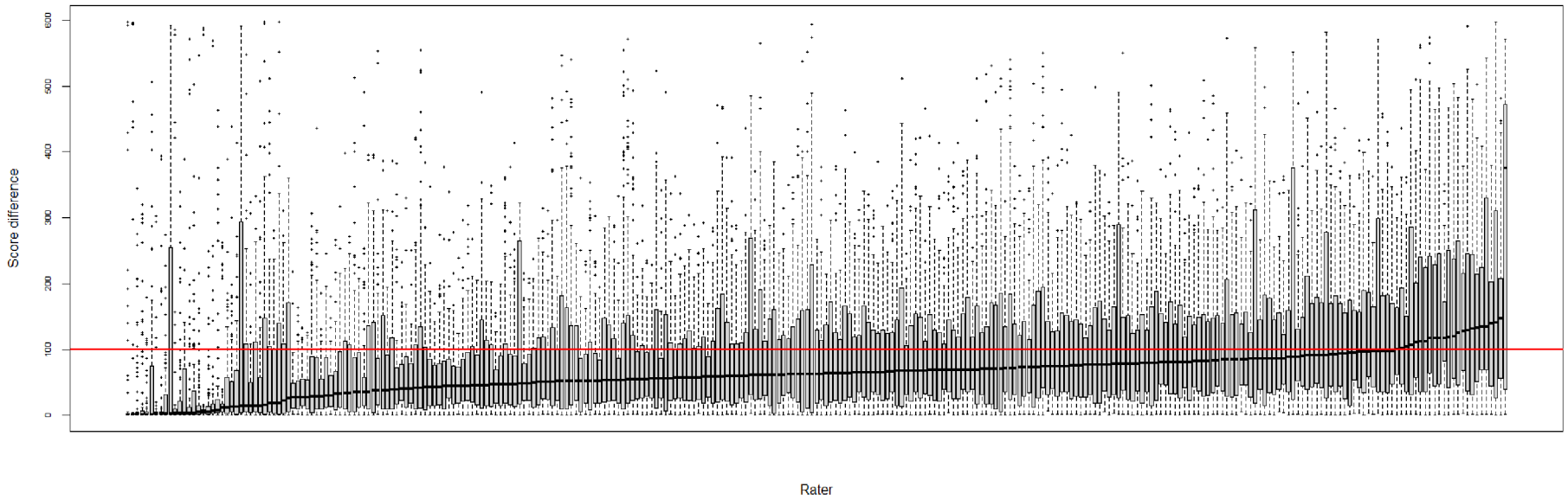


FIGURE 30: Score differences per rater. The black horizontal bars in the boxplots signify median differences between ratings for the same sample in two rating blocks. For most raters, the median difference is below 100 (out of 600 possible); a minority had a median difference between 100 and 150. The right-most rater who stands out with the median difference of 385.5 was removed from the data set.

or greater were discarded. If the rater clicked the replay button, this was assumed to mean that the rater was paying attention to the procedure and actively listening to the recording, so the RT was not considered even if it exceeded 10s (see also Section 6.2.5 above).

7 General discussion and conclusions

In Chapter 4 we reported a series of analyses of foreign-accent rating data that were prepared in order to address some common challenges in foreign-accent rating experiments. The outcomes of the foreign accent research was then used in linguistically informed evaluations of automatic language identification and speaker recognition technologies.

Some of our results with respect to foreign accent are in line with previous published research. The role of native language in foreign accent ratings—presented in Section 4.5—was found not to be completely straightforward: it seems that non-native raters can evaluate foreign accents in ways qualitatively similar to native raters; most non-native listeners rate native talkers as more accented and non-native talkers as less accented than do native raters; and L2 listeners tend to rate non-native talkers with the same native language as less accented than do other raters, however, this does not necessarily apply to all language combinations. Slovak raters were found to rate Czech talkers as less accented than did Czech talkers. Similar "mismatches" have been reported in literature before (Wester and Mayo 2014) but we are not aware of nor do we offer a universal explanation to this phenomenon. It is most likely that the experience with native and non-native English accents played a role in this result, as all Slovak talkers in the data set and most Slovak raters were students of English in the Czech Republic so they were certainly not the best representatives of the Slovak population eligible to be raters of foreign accent in English.

Some of the results of the accent rating experiments are a little more novel: we found that the definition of the rating task can be used to direct raters' attention to distinguish native speakers more categorically from non-native speakers by instructing listeners to rate their confidence about the perception of a foreign accent rather than rating accent strength itself. Even if the ratings of confidence and of accent strength were found to be strongly correlated in Section 4.1.3, the principal difference between them became important in relation to some factors affecting accent ratings and also when it came to correlations accent ratings and ASR scores. *Accent familiarity* was found to be correlated with ratings of confidence about accents, but not with accent-strength ratings. It seems that confidence ratings are more rater-centric in that they inform us about the raters' ability to identify foreign accents and they produce more extreme, as if binary, ratings "How sure are you that the speaker has a foreign accent" seems to be translated by the rater as "Does the speaker have a foreign accent –

yes or no?" In this way confidence ratings can be correlated with raters' accent familiarity, while accent strength ratings are more tightly linked to the variation of accents themselves. In contrast, the answers to the question "How strong is the foreign accent of the speaker" were found to be more strongly correlated with ASR scores, probably because of what was just said about the connection of accent strength ratings and accent variation.

Another finding that is in line with previous research is that adverse listening conditions made it more difficult to differentiate between degrees of foreign accent: everybody received more average accent ratings in the recordings in "Phone" quality when compared to "Original" studio recordings (Section 4.2). Furthermore, it turned out that non-native talkers who in the "Original" recordings received accent ratings similar to those of native talkers also grouped with native talkers when it came to accent ratings in adverse listening conditions: their ratings were more average, that is "on the phone" they sounded more accented foreign accented, or the raters were less sure that the talkers do *not* have a foreign accent, than in recordings with a full bandwidth that provided raters with more information about segmental features in the speech of the non-natives.

Chapter 5 is dedicated to biometric speech technologies. It is likely more conceivable that the native language of the talker should influence the accuracy of language ID technologies, more specifically, that a strong foreign accent should decrease LID accuracy. The connection between automatic speaker identification and foreign accents is less obvious and in my opinion less predictable. Would Czech speakers with a heavy accent in English sound more similar to what they sound like in Czech (because they retain phonetic and phonological features of their first language) or would they sound more different because their heavy foreign accent alters their usual intonation and speaking rate? Before addressing this question, it turned out to be more important to identify recording quality as the main predictor of the accuracy of both automatic language identification and automatic speaker recognition.

Both Phonexia and SpeechBrain technologies were negatively affected by the "harsh" conditions of the "Phone" recordings. In the case of `spkrec` and `lang-id`, it was—most probably—mainly because of the mismatch of the technical characteristics of the technology training data (16 kHz, YouTube and CommonVoice data) and the "Phone" data in our data set (8 kHz, band-filtered, etc.). The negative effect of low audio quality was evident, however, even in the case of `SID4-XL4` and `LID-L4`, which are trained on 8 kHz data. Our results provide evidence that channel mismatches and low data quality are more challenging to biometric speech technologies than are the effects of native language and foreign accent. When the factor of channel was under control, though, it was possible to see the effects of these less important factors, too. The results in Chapter 5 show that native language and foreign accent ratings do not necessarily correlate with language identification scores but they can be predictors of speaker identification scores when recording quality is factored out.

Some things remain unexplained, though. Why did `LID-L4` perform so much better with Czech data than with English data in both channels (Section 5.1.2)? Why were both ASR systems so much better in the Czech-only trials than in the English-only trials when the channel was factored out (Section 5.2.5)? Could it have to do with nativeness and the foreign accent or did the reasons lie purely in the training data and the architecture of the technologies? Another option is that the reason for the above-average performance of ASR technologies with the Czech data is related to the fact that some of the English samples were selected from different recording sessions as it was sometimes impossible to find single sessions which contained all the four phrases without any hesitations, external noises, etc. The Czech samples were selected based on similar criteria, but the control was not so strict since the Czech samples were never intended to be used in an accent-rating experiment. And so more of the Czech samples were in fact same-session samples. Session-mismatch is known to influence ASR scores and in some evaluations, same-session trials are explicitly excluded (see Morrison and Enzinger 2016).

There are, of course, other things in play which we did not consider with the importance they would deserve. We only used read speech both in the foreign accent experiments and the speech technology evaluations. We did not make use of the use of the semi-spontaneous recordings which were originally created for the purpose of a multi-modal evaluation. The reason was mainly the time-consuming nature of data preparation and also the difficulty to recruit raters for FAR experiments which prevented us from creating and deploying a more realistic set of stimuli, and thus from collecting more realistic foreign accent ratings for data that speech technologies—at least in the case of Phonexia software—are primarily designed to work with spontaneous speech.

Another aspect of the same problem is that there was a low variability of the test data: only the same four phrases were used over and over again. This was certainly very challenging for raters in the FAR experiment, and it most likely also affected the speech technology tests (and made them less valid, I am afraid). After all, the ASR evaluation reported in Section 5.2, resembles text-dependent speaker verification rather than text-independent speaker recognition which is the primary purpose of `SID4-XL4`. It is questionable how much the content of the phrases influenced the results but since the phrases were identical for *all* talkers, at least the speech technologies were presumably influenced in the same way for all talkers.

Another related problem was that when the data set of FA ratings and of the corresponding LID and ASR measurements was split up into groups according to several factors—native language of the talkers, native language of the raters, two versions of the rating question, two versions of recording quality—we ended up with little data in some of the groups. This is, of course, connected with the problem of recruiting experiment participants, with no funding and with a pandemic booming.

While evaluating the drawbacks of the presented research, what other possible flaws can be found in the experiment design or procedure? We mentioned in the

beginning of this chapter that the selection of Slovak talkers and raters was not methodologically quite "kosher". It should be noted that the decision to include a single talker with a native language (Ukrainian/Russian) different from the majority was admittedly controversial, too. Schmid and Hopp (2014) found that removing the most strongly accented L2 talkers increased foreign accent ratings of the whole L2 group with respect to native talkers, which, in their view, complemented the findings of Flege and Fletcher (1992), showing that removing some (or all) native talkers from rating experiments led to lower (i.e., more native-like) ratings for L2 talkers. However, at the time of the preparation of the accent rating experiment, I was overwhelmed by a "the-more-the-merrier" attitude and did not think through all the consequences.

Some studies have found ceiling effects in FA rating experiments (e.g. Southwood and Flege 1999). I have identified a similar effect which I am not sure if it is a flaw in experiment design or a valuable contribution to linguistics, and that is the fact that some raters used the continuous scale in combination with the task to rate their confidence of a perceived foreign accent not so much as a continuous scale but rather as a binary toggle switch: "Yes, speaker has a foreign accent" – "No, speaker does not have a foreign accent".

In the feedback to the rating experiment, some raters commented on the use of foreignness and nativeness in the formulation of the rating task and in the description of the rating scale. One U.K. rater interpreted "foreignness" as meaning "from a different country" even if the talker would be a native English speaker from Canada or Australia and conversely, he or she also considered the "foreign" category *not* to include native speakers of, say Welsh, if they did not speak English as native speakers but lived in the U.K. Other raters struggled with the fact that the scale contained both reference to nativeness and foreignness and they would have preferred to have something like "is a native speaker" vs "is not a native speaker". In summary, it may have been a mistake not to inform raters about some of the concepts that the experiment took for granted.

What were some of the more useful findings of the experiments? The effect of native language and foreign accent on the accuracy of biometric speech technologies in our data set was weak in comparison to the effect of recording quality. If the channel was factored out, then native language and foreign accent ratings could take effect and we could confirm that they should not be left out of consideration by technology users. To rephrase our summarizing assumption from Section 3.1.1: automatic speaker recognition and language identification systems have the *potential* of being language independent, but "we are not there yet". Matched-language trials in ASR are more likely to return high scores and therefore may be prone to false positive outcomes. In contrast, cross-language trials are more likely to return below-average scores and thus may be prone to false negative outcomes.

Similar conclusions can be drawn for the channel-mismatch problem: the challenging "Phone-only" matched-channel condition resulted in more false positives, while the cross-channel condition "Original-Phone" produced considerably more false negatives.

Both partial errors naturally contributed to higher Equal Error Rate values and thus the accuracy of the ASR system as a whole.

To conclude, when it comes to the effect of channel and native language on ASR, it turns out that the mismatch is not the biggest problem by itself. A bigger problem was when "all" channel or language combinations are taken together, as this results in a kind of "super mismatch" in that the matched-channel and cross-channel or alternatively matched-language and cross-language score distributions extend beyond the individual matched and mismatched score distributions and make room for false positive and false negative outcomes to thrive. Hopefully the findings of this thesis can contribute to curbing the false decisions and boosting the true ones.

Bibliography

Anderson-Hsieh, Janet, and Kenneth Koehler. 1988. "The effects of foreign accent and speaking rate on native speaker comprehension." *Language Learning* 38 (4): 561–613.

Aoyama, Katsura, and Susan S. Guion. 2007. "Prosody in second language acquisition. Acoustic analysis of duration and F0 range." In *Language experience in second language speech learning,* edited by Ocke-Schwen Bohn and Murray J. Munro. Amsterdam: Benjamins.

Baayen, R. H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Edmonton: Cambridge University Press.

Behrens, Susan J., and Sheila E. Blumstein. 1988. "Acoustic characteristics of English voiceless fricatives: a descriptive analysis." *Journal of Phonetics,* no. 16: 295–298.

Bent, Tessa, and Ann R. Bradlow. 2003. "The Interlanguage Speech Intelligibility Benefit." *The Journal of the Acoustical Society of America* 114 (3): 1600–1610.

Boersma, Paul, and David Weenink. 2018. *Praat: doing phonetics by computer [Computer program]. Version 6.1.04.* `http://www.praat.org`.

Bongaerts, Theo, Brigitte Planken, and Erik Schils. 1995. "Can Late Learners Attain a Native Accent in a Foreign Language? A Test of the Critical Period Hypothesis." In *The Age Factor in Second Language Acquisition,* edited by D. Singleton and Z. Lengyel, 30–50. Clevedon UK: Multilingual Matters.

Brownlee, Jason. 2020. *Failure of Classification Accuracy for Imbalanced Class Distributions.* `https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions`.

Bulejcik *v* R. 1996. "HCA 50. (1996) 185 CLR 375." `http://www8.austlii.edu.au/cgi-bin/viewdoc/au/cases/cth/HCA/1996/50.html`.

Český statistický úřad. 2020. *Terciární vzdělávání: Studenti a absolventi vysokoškolského a vyššího odborného vzdělávání.* Praha. `https://www.czso.cz/documents/10180/122323898/23006020p.pdf`.

Chung, J. S., A. Nagrani, and A. Zisserman. 2018. "VoxCeleb2: Deep Speaker Recognition." In *INTERSPEECH.*

Corder, Gregory W., and Dale I. Foreman. 2009. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach.* New Jersey: Wiley.

Corretge, Ramon. 2019. *Praat Vocal Toolkit [Computer program].* `http://www.praatvocaltoolkit.com`.

De Jong, Nivja H., and Ton Wempe. 2009. *Praat script to detect syllable nuclei and measure speech rate automatically.* `https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2`.

Derrick, B., A. Ruck, D. Toher, and P. White. 2018. "Tests for equality of variances between two samples which contain both paired observations and independent observations." *Journal of Applied Quantitative Methods* 13 (2): 36–47.

Desplanques, B., J. Thienpondt, and K. Demuynck. 2020. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification." In *Interspeech 2020,* edited by Helen Meng, Bo Xu, and Thomas Fang Zheng, 3830–3834. ISCA.

Doddington, George R., Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. 2000. "The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective." *Speech Communication,* no. 31: 225–254.

Drygajlo, Andrzej, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen, and Tuija Niemi. 2015. *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition.* ENFSI.

Edmond, Gary, Kristy Martire, and Mehera San Roque. 2011. "'Mere guesswork': Cross-Lingual Voice Comparisons and the Jury." *Sydney Law Review,* no. 33: 395–425.

Edwards, Jette G. Hansen, and Mary L. Zampini, eds. 2008. *Phonology and Second Language Acquisition.* Amsterdam: Benjamins.

Enzinger, Ewald, Geoffrey Stewart Morrison, and Felipe Ochoa. 2016. "A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case." *Science and Justice,* no. 56: 42–57.

Flege, J. E., and J. K. Fletcher. 1992. "Talker and listener effects on degree of perceived foreign accent." *Journal of the Acoustical Society of America,* no. 91: 370–389.

Flege, James Emil. 1984. "The detection of French accent by American listeners." *Journal of the Acoustical Society of America* 76 (3): 692–707.

———. 1988. "Factors affecting degree of perceived foreign accent in English sentences." *Journal of the Acoustical Society of America* 84 (1): 70–79. https://pdfs.semanticscholar.org/7b96/35e4478d2ac0f6041ad7b24cc8f6c378459f.pd.

Flege, James Emil, Elaina M. Frieda, and Takeshi Nozawa. 1997. "Amount of native-language (L1) use affects the pronunciation of an L2." *Journal of Phonetics* 25 (2): 169–186.

Grabe, Esther, and Ee Ling Low. 2002. "Durational variability in speech and the rhythm class." Edited by C. Gussenhoven and N. Warner, (Berlin): 515–546.

Hanulíková, Adriana, and Silke Hamann. 2010. "Illustrations of the IPA: Slovak." *Journal of the International Phonetic Association* 40 (03): 373–378. DOI:10.1017/S0025100310000162.

Huang, Becky H. 2013. "The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech." *System,* no. 41: 770–785. https://www.sciencedirect.com/sdfe/reader/pii/S0346251X13000973/pdf.

Institute for Telecommunication Sciences. 1996. "Definition: voice frequency (VF)." https://www.its.bldrdoc.gov/fs-1037/dir-039/_5829.htm.

International Telecommunications Union. 2006. *G.723.1 : Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s.* https://www.itu.int/rec/T-REC-G.723.1-200605-I/en.

Jamieson, Susan. 2004. "Likert scales: how to (ab)use them." *Medical Education* 38 (12): 1217–1218.

Jesney, Karen. 2004. *The Use of Global Foreign Accent Rating in Studies of L2 Acquisition.* Calgary: AB: Language Research Centre, University of Calgary. https://bcf.usc.edu/~jesney/Jesney2004GlobalAccent.pdf.

Jessen, Michael, Jakub Bortlík, Petr Schwarz, and Yosef A. Solewicz. 2019. "Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (*forensic_ eval_ 01*)." *Speech Communication,* no. 111: 22–28. https://doi.org/10.1016/j.specom.2019.05.002.

Kerstholt, José, Noortje J. M. Jansen, Adri G. Vanamelsvoort, and A. P. A. Broeders. 2004. "Earwitnesses: Effects of Speech Duration, Retention Interval and Acoustic Environment." *Applied Cognitive Psychology,* no. 18: 327–336.

Künzel, Herman, and Paul Alexander. 2014. "Forensic Automatic Speaker Recognition with Degraded and Enhanced Speech." *Journal of the Audio Engineering Society* 62 (4): 244–253. https://doi.org/10.17743/jaes.2014.0014.

Kuschmann, Anja, and Anja Lowit. 2015. "The role of speaking styles in assessing intonation in foreign accent syndrome." *International Journal of Speech-Language* Pathology 17 (5): 489–499. https://strathprints.strath.ac.uk/50689/6/Kuschmann_Lowitt_IJSLP_2015_Role_of_speaking_styles_in_assessing_intonation.pdf.

lang-id-commonlanguage_ecapa. 2021. *Language Identification from Speech Recordings with ECAPA embeddings on CommonLanguage.* https://huggingface.co/speechbrain/lang-id-commonlanguage_ecapa.

Lecumberri, Maria Luisa Garcia, Martin Cooke, and Anne Cutler. 2010. "Non-native speech perception in adverse conditions: A review." *Speech Communication,* no. 52: 864–886.

Llisterri, Joaquim. 1992. "Speaking styles in speech research." In *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language. Dublin, Ireland, 15-17 July 1992.* http://liceu.uab.es/~joaquim/publicacions/SpeakingStyles_92.pdf.

Marxer, Ricard, Jon Barkerb, Najwa Alghamdib, and Steve Maddockb. 2018. "The impact of the Lombard effect on audio and visual speech recognition systems." *Speech Communication,* no. 100: 58–68.

Monson, Shaun. 2005. *Earthlings – 10 Year Anniversary Edition [Documentary].* https://vimeo.com/209647801.

Morrison, Geoffrey Stewart, and Ewald Enzinger. 2016. "Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (*forensic_ eval_ 01*) – introduction." *Speech Communication,* no. 85: 119–126. http://dx.doi.org/10.1016/j.specom.2016.07.00.

Morrison, Geoffrey Stewart, Felipe Ochoa, and Tharmarajah Thiruvaran. 2012. "Database selection for forensic voice comparison." In *Proceedings of Odyssey 2012: The Speaker and Language Recognition Workshop,* 62–77. Singapore.

Moyer, Alene. 2013. *Foreign Accent.* Cambridge: Cambridge University Press.

Munro, Murray J., and Tracey M. Derwing. 1995. "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners." *Language Learning,* no. 49 (Supp. 1): 285–310.

———. 1998. "The Effects of Speaking Rate on Listener Evaluations of Native and Foreign-Accented Speech." *Language Learning* 48 (2): 159–182.

———. 2001. "Modeling perceptions of the accentedness and comprehensibility of l2 speech: the role of speaking rate." *Studies in Second Language Acquisition,* no. 23: 451–468.

Munro, Murray J., Tracey M. Derwing, and Susan L. Morton. 2006. "The mutual intelligibility of L2 speech." *Studies in Second Language Acquisition* 28 (1): 111–131.

Nagrani, A., J. S. Chung, and A. Zisserman. 2017. "VoxCeleb: a large-scale speaker identification dataset." In *INTERSPEECH.*

Ng, Cheng Teng, Liya E Yu, Choon Nam Ong, Boon Huat Bay, and Gyeong Hun Baeg. 2018. "The use of Drosophila melanogaster as a model organism to study immune-nanotoxicity." *Nanotoxicology* 13 (4): 429–446. doi:10.1080/17435390.2018.1546413.

Olson, Linda L., and S. Jay Samuels. 1973. "Relationship between age and accuracy of foreign language pronunciation." *Journal of Educational Research* 66 (6): 263–268.

Omniglot. 2021. *English language, alphabet and pronunciation.* https://www.omniglot.com/writing/english.htm.

Oxenham, Andrew J. 2012. "Pitch Perception." *Journal of Neuroscience* 32 (39): 13335–13338. https://doi.org/10.1523/JNEUROSCI.3815-12.2012.

comprehensibility. 2021. In *OLD Online.* Oxford University Press, January, by Oxford Learner's Dictionary. Accessed January 8, 2021. https://www.oxfordlearnersdictionaries.com/definition/english/comprehensibility.

Pellegrino, Elisa. 2012. "The perception of foreign-accented speech. Segmental and suprasegmental features affecting the degree of foreign accent in L2 Italian." In *Proceedings of the VIIth GSCP International Conference: Speech and Corpora,* edited by Heliana Mello, Massimo Pettorino, and Tommaso Raso.

Perrachione, Tyler K. 2019. "Recognizing speakers across languages." In *The Oxford Handbook of Voice Perception,* edited by Sascha Frühholz and Pascal Belin. Oxford: Oxford University Press.

R Core Team. 2021. *R: A language and environment for statistical computing.* Vienna, Austria.

Ravanelli, Mirco, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, et al. 2021. *SpeechBrain: A General-Purpose Speech Toolkit.* ArXiv:2106.04624. arXiv: 2106.04624 [eess.AS].

Sato, Charlene J. 1991. "Sociolinguistic variation and language attitudes in Hawaii." In *English around the world: Sociolinguistic perspectives,* edited by J. Cheshire, 647–663. Cambridge: Cambridge University Press.

Schmid, Monika S., and Holger Hopp. 2014. "Comparing foreign accent in L1 attrition and L2 acquisition: Range and rater effects." *Language Testing* 31 (3): 367–388.

Šimáčková, Šárka, Václav Podlipský, and Kateřina Chládková. 2012. "Czech spoken in Bohemia and Moravia." *Journal of the International Phonetic Association* 42 (2): 225–232.

Šimáčková, Šárka, Václav Jonáš Podlipský, and Kateřina Kolářová. 2014. "Linking Versus Glottalization: (Dis)connectedness of Czech-Accented English." In *Proceedings of the International Symposium on the Acquisition of Second Language Speech,* 678–692.

Sinisetty, Ganesh, Pavlo Ruban, Oleksandr Dymov, and Mirco Ravanelli. 2021. *CommonLanguage (0.1) [Data set].* `https://doi.org/10.5281/zenodo.5036977`.

Smith, Harriet M. J., Thom S. Baguley, Jeremy Robson, Andrew K. Dunn, and Paula C. Stacey. 2018. "Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance." *Applied Cognitive Psychology,* no. 33: 272–287.

Smith, Julius O. 2007. *Mathematics of the Discrete Fourier Transform (DFT), with Audio Applications.* `https://ccrma.stanford.edu/~jos/mdft/Pulse_Code_Modulation_PCM.html`.

Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. "X-vectors: Robust DNN embeddings for speaker recognition." `https://www.danielpovey.com/files/2018_icassp_xvectors.pdf`.

Southwood, M. Helen, and James E. Flege. 1999. "Scaling foreign accent: direct magnitude estimation versus interval scaling." *Clinical Linguistics and Phonetics* 13 (5): 335–349.

Spilková, Helena, and Wim A. van Dommelen. 2010. "English 'of' in L1 and L2 speakers' read and spontaneous speech," no. 54: 91–96. `https://www.researchgate.net/publication/266047717`.

spkrec-ecapa-voxceleb. 2021. *Speaker Verification with ECAPA-TDNN embeddings on Voxceleb.* `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`.

Stocker, Ladina. 2017. "The Impact of Foreign Accent on Credibility: An Analysis of Cognitive Statement Ratings in a Swiss Context." *Journal of Psycholinguistic Research,* no. 46: 617–628.

Stoet, Gijsbert. 2010. "PsyToolkit – A software package for programming psychological experiments using Linux." *Behavior Research Methods* 42 (4): 1096–1104.

———. 2017. "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments." *Teaching of Psychology* 44 (1): 24–31.

Trofimovich, Pavel, and Wendy Baker. 2006. "Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech." *Studies in Second Language Acquisition* 28 (1): 1–30. https://doi.org/10.1017/S0272263106060013.

Ulbrich, Christiane, and Ineke Mennen. 2015. "When prosody kicks in: The intricate interplay between segments and prosody in perceptions of foreign accent." *International Journal of Bilingualism:* 1–28. DOI:10.1177/1367006915572383.

Vel. 2017. *Masters/Doctoral Thesis (LaTeX template).* https://www.latextemplates.com/template/masters-doctoral-thesis.

Volín, Jan, and Radek Skarnitzl. 2010. "The strength of foreign accent in Czech English under adverse listening conditions." *Speech Communication,* no. 52: 1010–1021.

Wester, Mirjam, and Cassie Mayo. 2014. "Accent rating by native and non-native listeners." In *Proceedings of ICASSP,* 7749–7753. Florence, Italy. http://www.cstr.ed.ac.uk/downloads/publications/2014/wester:icassp:14.pdf.

Worldatlas. 2021. *The Most Spoken Languages in America.* https://www.worldatlas.com/articles/the-most-spoken-languages-in-america.html.

Zimmermann, Jacqueline F., Morris Moscovitch, and Claude Alaina. 2016. "Attending to auditory memory." *Brain research,* no. 1640: 208–221.

Shrnutí

Tato dizertační práce se v sedmi kapitolách zabývá tématem českého přízvuku v angličtině z pohledu lingvisticky a biometrických řečových technologií. První kapitola stručně uvádí do problematiky, popisuje motivaci pro tuto práci a určuje její rámec. Ve druhé kapitole práce shrnuje poznatky výzkumu věnovaného hodnocení cizího přízvuku. Představuje základní principy a především praktické otázky metodologie výzkumu cizího přízvuku s cílem vytyčit zásady pro realizaci vlastního výzkumu. Jde především o otázky shody rodného jazyka mezi hodnotiteli a mluvčími, znalosti hodnoceného přízvuku, formulace úkolu hodnocení a jeho vlivu na hodnotitele, řečového stylu a hodnocení přízvuku ve ztížených poslechových podmínkách. Ve druhé kapitole jsou také identifikovány některé problematické aspekty stávající výzkumné praxe v tomto oboru, jako je například používání Likertovy škály pro získávání hodnocení přízvuku a jejich následné zpracování nevhodnými statistickými metodami.

Třetí kapitola uvádí do problematiky biometrických řečových technologií, především s ohledem na aspekty, které souvisejí s tématem cizího přízvuku. V kapitole jsou nejprve představeny rozdílné požadavky kladené na systémy automatického rozpoznání řečníka (ASR, podle anglického "automatic speaker recognition") a rozpoznávání řečníků prováděné lidskými posluchači, například ve forenzní praxi. Kapitola dále popisuje výzvy, se kterými se obor rozpoznání řečníka typicky potýká, jmenovitě jde o neshodu v tzv. kanále (čili akustických charakteristikách nahrávek způsobených jejich zdrojem či přenosem) a neshodu v jazyku nahrávek, pro které se rozpoznání řečníka provádí. Dále jsou představeny základní typy úkolů řešených systémy pro ASR: identifikace a verifikace, které se odlišují počtem prováděných porovnání a hlavně teoretickým předpokladem o známosti či neznámosti řečníků, jejichž hlas se porovnává. V souvislosti s těmito úkoly ASR jsou popsány základní metody vyhodnocování přesnosti takových systémů. Oddíl věnovaný rozpoznání řečníka je zakončený popisem dvou systému ASR, komerčního systémů `SID4` firmy Phonexia a open-source technologie `spkrec-ecapa-voxceleb`. Poslední oddíl třetí kapitoly tvoří stručný úvod do problematiky automatického rozpoznání jazyka (LID, podle anglického "language identification"), a vyhodnocování přesnosti LID technologií. A v závěru jsou představeny dvě takové technologie na rozpoznání jazyka: `LID-L4` firmy Phonexia a open-source knihovna `lang-id-commonlanguage_ecapa` z projektu SpeechBrain.

Čtvrtá kapitola stanovuje konkrétní výzkumné otázky, hypotézy a predikce týkající se hodnocení přízvuku, a představuje experimenty provedené pro ověření těchto predikcí. Prvním dílčím tématem čtvrté kapitoly je formulace úkolu hodnocení cizího přízvuku a způsob, jakým tato formulace ovlivňuje pozornost hodnotitele a následně rozsah hodnot nasbíraných pro hodnocené nahrávky. Druhým tématem je hodnocení přízvuku ve ztížených poslechových podmínkách, kde se dávají do kontrastu originální studiové nahrávky a hodnocení přízvuku pro ně s hodnoceními přízvuku pro nahrávky upravené tak, aby zněly jako telefonní data. Třetím tématem je artikulační tempo, jeho rozdíl v řeči rodilých a nerodilých mluvčích a jeho vliv na hodnocení přízvuku. Dále pak kapitola analyzuje znalost rodného a cizího přízvuku a jakožto možné faktory ovlivňující hodnocení přízvuku. Nakonec se kapitola věnuje problematice shody rodného jazyka u mluvčího a hodnotitele. Práce dochází k závěru, že formulaci úkolu hodnocení přízvuku lze využít k nasměrování pozornosti hodnotitele buď k hodnocení vlastní *jistoty*, zda hodnocený mluvčí má, nebo nemá cizí přízvuk, případně k nasměrování pozornosti na samotnou *sílu* vnímaného cizího přízvuku. Dalším zjištěním je, že nerodilí mluvčí angličtiny, kteří získali podobné hodnocení přízvuku v podmínkách původních nahrávek jako rodilí mluvčí, jsou v podmínkách telefonních nahrávek hodnoceni taky podobným způsobem jako rodilí mluvčí, tedy jakoby měli silnější cizí přízvuk. Naproti tomu méně pokročilí mluvčí jsou hodnoceni tak, že mají silnější přízvuk v podmínkách studivých nahrávek, a v podmínkách simulujících telefonní hovor jako že mají přízvuk slabší. Dále docházíme k závěru, že míra znalosti rodilého přízvuku v angličtině (nebo spíše rodilých přízvuků) může být prediktorem hodnocení jistoty, že posluchač slyší cizí přízvuk, ale nezdá se, že by byla prediktorem hodnocení síly přízvuku. Kapitolu zakončuje potvrzení již dříve známého poznatku, že shoda rodného jazyka mezi hodnotitelem a mluvčím má vliv na snížení průměrných hodnocení přízvuku pro konkrétní nahrávky, nicméně nezaručuje, že hodnotitelé s jiným rodným jazykem nebudou dané nahrávky hodnotit jako by měly ještě menší míru přízvuku.

Pátá kapitola, věnovaná otázkám řečových technologií v souvislosti s hodnocením cizího přízvuku, identifikuje tvz. kanál nahrávky jako hlavní prediktor přesnosti technologií pro automatické rozpoznání řečníka a jazyka s tím, že všechny analyzované technologie fungují lépe na originálních studiových nahrávkách a jejich přesnost se někdy zásadním způsobem snižuje, pokud jsou použity na nahrávky simulující kvalitu telefonních dat. Kapitola ukazuje, že rodný jazyk mluvčího a hodnocení cizího přízvuku pro jeho nahrávky nutně nekorelují s přesností automatického rozpoznání jazyka. Minimálně v případě obou použitých systémů na rozpoznání jazyka jsou výsledky analýzy vlivu rodného jazyka zásadně odlišné: zatímco systém `LID-L4` je podstatně přesnější s nahrávkami v rodném českém jazyce než s nahrávkami v cizím jazyce, je to u systému `lang-id` spíše naopak. Naproti tomu v případně automatického rozpoznání řečníka je rodný jazyk mluvčích a hodnocení cizího přízvuku pro anglické nahrávky poměrně silně korelován s hodnocením cizího přízvuku, nicméně

především až po té, co je odstraněn dominující efekt kanálu. Představená analýza upozorňuje na úskalí přílišného rozdělování dat podle velkého množství prediktorů, které může vést k přílišnému snížení počtu vzorků ztěžujícímu statistické vyhodnocení dat.

Šestá kapitola doplňuje metodologické sekce z jednotlivých podkapitol o detailní popis metody použité pro vytvoření vícejazyčného datasetu, obsahujícího nahrávky čtyř rodilých mluvčích angličtiny, 32 rodilých mluvčích češtiny, pěti rodilých mluvčích slovenštiny a jedné ukrajinsko-ruské rodilé mluvčí. Kapitola podává přehled metainformací k datasetu a následně popisuje proces, kterým byla část datasetu zpracována pro potřeby experimentu na hodnocení cizího přízvuku. Následuje představení samotného online experimentu, přehled jeho účastníků – hodnotitelů přízvuku, a popis procesu následného zpracování dat. Závěrečná sedmá kapitola nabízí shrnutí poznatků, ke kterým práce v jednotlivých kapitolách došla a nastiňuje případně nezodpovězené nebo nově položené otázky.