



UNIVERSITÀ DEGLI STUDI DI TRENTO
CIMeC - Center for Mind/Brain Sciences

Master's Degree in Cognitive Science

Academic Year: 2018-2019

A deep learning approach to spoken proficiency scoring for non-native speakers

STUDENT: Ms. Nina Hosseini Kivanani

SUPERVISOR: Dr. Carlo Strapparava

CO-SUPERVISOR: Dr. Tommaso Caselli

A deep learning approach to spoken proficiency scoring for non-native speakers

Nina Hosseini Kivanani: MSc's Thesis

Erasmus Mundus Joint Master's Degree in Language & Communication

Technologies

University of Groningen

University of Trento

Supervisor: Dr. Carlo Strapparava

Co-supervisor: Dr. Tommaso Caselli



UNIVERSITÀ DEGLI STUDI DI TRENTO

CIMeC - Center for Mind/Brain Sciences



university of
 groningen



FONDAZIONE
BRUNO KESSLER



With the support of the
Erasmus+ Programme
of the European Union



Acknowledgment

This thesis would not have been possible without the support of my advisors, friends, and family. Foremost, I would like to thank the SpeechTek lab in FBK that helped me through this journey, and I would like to give a particular thanks to Roberto Gretter, Daniele Falavigna, and Marco Matassoni. Thanks for all the coffee times for discussing my thesis progress and supporting through learning bash scripting and giving me enough freedom to learn from my own mistakes. They are my mentors. Secondly, I would like to thank my supervisors: Dr. Carlo Strapparava from Trento University and Dr. Tommaso Caselli from Groningen University. Spending two years of my life in Groningen and Trento, gave me a lot of useful and valuable experiences. It's difficult to say goodbye to the beautiful and friendly city of Trento after my second year of studies. I can say that my good memories come from Trento and my work with FBK. This two-year journey was made possible by a scholarship (the Erasmus Mundus scholarship from the European Union) to study in the Erasmus Mundus European Master Program in Language and Communication Technologies (EM LCT), and this thesis is the output of that fabulous opportunity. Last but not least, thanks to my family, friends (Svetlana, Simon, Aria, Homa, and Gosse), and my colleagues in SpeechTek for supporting me every day. To all of the above and to anyone I missed who has touched my life and brought me to this point: **Thank you.**

Abstract

Speech assessment is one of the key tasks in Computer Assisted Pronunciation Training (CAPT). In this study, we develop a mispronunciation assessment system that checks the pronunciation of non-native English speakers and identifies which phonemes were pronounced incorrectly. Thus, this work mainly focuses on automatic estimation of the common mispronounced phonemes of Italian learners of English, both adults, and children, and presents an analysis and evaluation of the native and non-native pronunciation observed in phonetically annotated speech corpora. For this, we design an Automatic Speech Recognition (ASR) system that evaluates learners' speech at the phoneme level to detect pronunciation errors. For mispronunciation assessment, we select two corpora (Interactive Spoken Language Education (ISLE) and ChildEn), create a new list of phonemes extracted from manually annotated data from both corpora, then train an acoustic model with the new phoneme set. These evaluations rely on ASR, which could be performed using a database of non-native speakers of frequent mistakes.

In detail, two language models were considered; an n-gram model and an error language model. The output of the n-gram model helped us obtain insights into the phone errors in our audio files. Applying the error language model is the novel approach to the second language (L2) speech assessment, which combines both true transcriptions for a word with their adapted transcriptions for the same word. Quantitative evaluations are carried out for Italian learners of English. The output of the preliminary results reveals that Italian learners of English have the most problems in producing certain phonemes that do not appear in their language. Our results show that the selected error model can discriminate correct sounds from incorrect sounds in both native and nonnative speech, and therefore can be used to detect pronunciation errors in non-native speech. The Phone error rates show improvement in using the error language model. In summary, our ASR system shows high accuracy after applying the error language model on our selected corpora.

Contents

Acknowledgement	iii
Abstract	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Statement of problem	1
1.2 Purpose of study	2
1.3 Significance of the study	3
1.4 Research Context & Support	4
1.5 Outline of the thesis	4
2 Related work	5
2.1 ASR for error detection	5
2.1.1 Holistic pronunciation error detection	6
2.1.2 Pinpoint pronunciation error detection	7
3 Description of data	9
3.1 Speech corpora	9
3.1.1 ChildEn corpus	9
3.1.2 ISLE corpus	10
3.2 Choosing a phone set	11
3.3 Phone errors: ISLE & ChildEn	13
3.3.1 Corpora statistics	15
4 Material & Method	16
4.1 Speech file transcriptions: Ideal, Manual, & ASR	16
4.1.1 Sample	18
4.2 Basic framework of speech analysis	19
4.2.1 ASR evaluation	20
4.2.2 Language model: N-Grams	21
4.2.3 Language model: Error model	22

4.2.4	Acoustic model	24
5	Results	26
5.1	Detection output	26
5.1.1	N-Grams output	26
5.1.2	Error model output	28
5.2	Evaluation	29
5.3	Summary	31
6	Final discussion & Conclusion	32
6.1	Directions for future research	34
6.1.1	Feedback	34
6.1.2	Duration	35
	Appendices	36
	Appendix A List of Abbreviations	36
	Appendix B Words & Phones list	37
	Appendix C PER: N-Gram output	41
	Appendix D Sample output of 5-gram model: ISLE & ChildEn	43
	Appendix E Adapted lexicon	46
	Appendix F Phone list of Ideal, Manual, & ASR	47
	Bibliography	52

List of Figures

1	Basic nature of Speech Recognition	3
2	Vowel charts of English (left graph) & Italian (Right graph).	12
3	Frequency of degree of “goodness” of pronunciation	13
4	Phoneme errors distributions of ISLE (left graph) & ChildEn (Right graph).	15
5	A TextGrid of manual annotation example from ChildEn corpus.	18
6	Components and models for ASR	19
7	Pronunciation error types by categories (Witt, 2012).	22
8	Overall detection accuracy for ISLE (left graph) & ChildEn (Right graph).	30
9	Phoneme distributions of ISLE (left graph) & ChildEn (Right graph).	40

List of Tables

1	Information regarding the selected corpora	10
2	Statistics about the speech data sets used for ASR in this study	11
3	Language characteristics of English & Italian	12
4	Information of words & phone errors in ISLE & ChildEn	13
5	Most frequent words in ISLE & ChildEn	14
6	Sample transcription: Manually, Ideally, & ASR	17
7	Error rules: Substitutions, Deletions, & Insertions	23
8	Sample of adapted pronunciations for the words “Autumn” & “Happy”	24
9	Example of forced alignment result from Kaldi toolkit: A birthday cake	25
10	PER output of n-gram models	27
11	Output of number of errors & PER based on 5-gram model	28
12	Output of PER based on error model	29
13	Detection Accuracy: n-gram & error model	30
14	Examples of English reference transcriptions & the selected transcription	33
15	Phoneme list extracted from MAN with examples & their transcriptions	37
16	Common words in both ISLE & ChildEn	39
17	PER of 1-gram model	41
18	PER of 2-gram model	41
19	PER of 3-gram model	42
20	PER of 4-gram model	42
21	PER of 5-gram model	42
22	Output of top frequent errors for ISLE corpus generated form 5-gram model	44
23	Output of top frequent errors for ChildEn corpus generated from 5-gram model	45
24	Sample of the adapted lexicon	46
25	ISLE phone list: Ideal, Manual & ASR	47
26	ChildEn phone list: Ideal, Manual & ASR	50

1 Introduction

1.1 Statement of problem

The number of people who are learning a second language (L2) worldwide is increasing. Consequently, the need to evaluate and grade their pronunciation is becoming an important topic. [Fant \(1973\)](#) defines mispronunciation as surface pronunciation forms differing from canonical pronunciation forms. **Phoneme level mispronunciation** refers to the interference of a second language learner’s native language during speech production, where foreign sounds are produced in a way similar to a phoneme in their native language. The persistent presence of high error rates in speech recognition domains resulting from mispronunciations motivates us to attempt to find alternative techniques for handling mispronunciations ([Figure 1](#)).

Based on [Huensch \(2019\)](#), in the second language (L2) learning progress, pronunciation plays an important role, although this part of the learning process has always received less attention due to a lack of time and resources compared to other skills in classrooms ([McCrocklin, 2016](#)). Pronunciation often gets ignored when time constraints force teachers to make choices about what they can cover during class time. Students rarely get the number of pronunciation instructions that they need or want. Pronunciation is one of the fundamentals of language learning, and it is considered a primary factor of spoken language when it comes to an understanding and being understood by others. Most of the pronunciation activities in classrooms rely on the teacher to monitor, evaluate, and provide feedback on student pronunciation. This traditional technique does not seem adequate to correct student pronunciation, and it tends to be costly and time-consuming.

Given these constraints, the growing tendency to assess non-native language leads to increased interest in automatic proficiency assessment of speech and boost Computer Assisted Language Learning (CALL) tools in the field of language teaching for L2 learners. CALL tools are designed to recognize words or sentences uttered by L2 learners by using an Automatic Speech Recognition (ASR) system. The computer-assisted pronunciation training (CAPT) system is one of the essential tools of CALL designed for automatically evaluating and detecting learner’s pronunciation errors.

In the CAPT system, pronunciation evaluation can happen at two levels: pinpoint error

detection (i.e., detecting specific pronunciation errors) and holistic detection (i.e., an overall assessment of a speaker’s proficiency) (Peabody, 2011). Apart from that, there are a number of well-known automatic pronunciation assessment tools for educational purposes for English learners such as the SpeechRater (Zechner et al., 2009) and the Versant system (Downey et al., 2008), which employ ASR systems to assess the goodness of pronunciation (GoP) (Witt and Young, 2000) to provide pronunciation evaluation from the phonemic to the prosodic domain (e.g., stress, rhythm, intonation). The main research questions of this study are as follows:

- Investigating two corpora (ISLE and ChildEn), which are the most frequent phoneme errors?
- Which kind of error models will be useful in detecting mispronunciation of L2 learners of Italian?
- Which speech materials will bring the highest accuracy for our detector model; (1) use of native or (2) use of non-native speech as training materials?

Ideally, our system should be able to detect errors the same as human annotators do. In other words, our system judgment should resemble human judgments. The selected corpora have a variety of contextualized speaking tasks other than read-aloud tasks that were manually annotated (more details in chapter 4).

1.2 Purpose of study

Fu et al. (2020) states three types of automatic pronunciation evaluation (for more details, see chapter 2: (i) Evaluation of a sentence or word, (ii) Evaluation of a specific phoneme, and (iii) Detection of mispronounced phonemes. This study seeks to examine the use of Automatic Speech Recognition (ASR), in order to empower learners to practice and improve their pronunciation on their own. The scope of this study is to develop an automatic grader based on extracted features related to the pronunciation of phonemes which can enhance the performance of an ASR system and provide feedback regarding pronunciation errors (mispronunciation) of learners on various time-scales including the segmental (i.e., minimal phonetic units: vowels and consonants), suprasegmental (i.e., stress and intonation), and the voice-setting aspect (i.e., voice position in sound articulation) (Fant, 1973). Among the mentioned time-scales, the suprasegmental aspect is less reliable than other aspects of pronunciation, making it more challenging to

reach a perfect model. Mastery of the segmental aspect provides an L2 learner with a native-like accent, and the suprasegmental aspect provides the learner with intelligibility (Raux and Kawahara, 2002; Derwing and Munro, 2005).

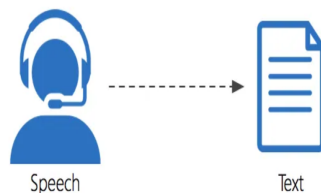


Figure 1: Basic nature of Speech Recognition

For this project, audio is passed through an automatic speech recognizer (ASR), and the recognized text is used with the audio to extract features (see Chapter 4). In more detail, the phonetic output that ASR produces could be used for error detection.

Previous research has focused on pronunciation grading during pronunciation assessment. For example, Gretter et al. (2019) developed a system to grade Italian L2 learners of German and English automatically, Wang et al. (2018) focus on monitoring the performance of the Educational Testing Service (ETS) SpeechRater tool.

The focus of this study is on pronunciation error detection. The language models that will be used for this study are n-gram models (see Chapter 5). Our acoustic model is based on a Hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) and is trained on speech data in a given data set.

1.3 Significance of the study

The findings of this study will help language teachers and improve autonomous learning for language learners in their pronunciation progress. For learners, the most challenging part of learning a new language is the pronunciation part, due to the fact that it is challenging to imitate sounds that are different to those of the native language’s phoneme inventory (Peabody, 2011). Using a well-designed ASR system allows students to work autonomously while also offering flexibility that can lead to the improvement of their pronunciation. Moreover, this helps students to direct their learning and bring any desired materials into practice with the program.

Besides, this study uses two corpora (ISLE and ChildEn) with different tasks from both children and adult Italian speakers that cover the diversity of speakers’ speech. These corpora

contain both spontaneous and read-aloud tasks. ASR systems perform well when the conditions are well controlled, which differs from the speech in noisy conditions, with different accents and different speech rates among speakers. To deal with this, we focus on feature extraction.

1.4 Research Context & Support

This thesis stems from a collaboration with a research institute, namely the Fondazione Bruno Kessler (FBK). The project was done in the SpeechTek lab ¹ lead by Giuseppe Daniele Falavigna at FBK. The research results of this study will be submitted to a upcoming conference. This study was supported by the Erasmus Mundus scholarship from the European Union as the double-degree program (Language and Communication Technology-LCT) ².

1.5 Outline of the thesis

The thesis is organized into the following five chapters: Chapter 2 will focus on the related studies on ASR. Chapter 3 describes the data and the preliminary results of error evaluation in this study. Chapter 4 presents the preparation of our database for evaluation of the proposed method focusing on the features considered for our machine learning approach. Chapter 5 covers the results of our ASR model based on the selected corpora. Chapter 6 provides a general discussion of the results, followed by concluding remarks.

Remarks: A list of abbreviations used throughout this study is given in Appendix A. Phonetic symbols used for notation are listed in Appendix B.

¹<https://ict.fbk.eu/units/speechtek/>

²<https://lct-master.org/>

2 Related work

This chapter details the literature related to detecting pronunciation errors (see Chapter 1). Error detection of mispronounced input is similar to the process in which an annotator manually indicates pronunciation errors. With respect to comparing the traditional way (teacher-based method) and using technology (ASR system) on improving pronunciation, [Kruk \(2012\)](#) reported that the experimental group (autonomy group) on using technology (computer-based work) performed better than the traditional group (traditional classroom controlled by teacher). This study tried to evaluate the autonomy in pronunciation learning.

Generally, the recent development of ASR systems, especially language learning, increases the opportunities for learners to receive feedback and autonomy practice to fill the gap in the classroom. The following sections describe research on holistic and pinpoint pronunciation error detection for background information. However, this study focuses primarily on pinpoint pronunciation detection.

2.1 ASR for error detection

In the field of automatic pronunciation assessment, various features were evaluated such as GoP, speaking rate, articulation rate, phonation time ratio (for more information see [Ryu et al. \(2017\)](#)); apart from these features, ETS presents 29 candidate features ¹ for scoring the speech of non-native learners of English. There are two common desired formats for automatic pronunciation assessment: unstructured spontaneous speech and read-aloud sentences. Unstructured spontaneous speech, in which the participants are asked questions that they must answer on the spot, and read-aloud sentences, in which the participants are provided a text to read and record. The more challenging format that received more attention is spontaneous speech at the level of multiple sentences ([Van Doremalen et al., 2009](#); [Hönig et al., 2010](#); [Nicolao et al., 2015](#)).

In addition to the desired formats, there are different approaches to assessing pronunciation, such as not using native speakers' data to evaluate speech ([Van Doremalen et al., 2009](#); [Hönig et al., 2010](#); [Maqsood et al., 2016](#)). As it was mentioned in the previous research (e.g., [Witt, 2012](#)), the importance of intelligibility of an utterance will compensate native speaker similarity, particularly for less advanced learners.

¹These features are related to the fluency or duration of silences of spoken English.

Prior work (e.g., [Amdal et al., 2009](#); [Strik et al., 2009](#)) mainly focused on specific phone pairs for mispronunciation detection. Four types of classifiers (Linear Discriminant Analysis (LDA), Linear Discriminant Analysis–Acoustic-Phonetic feature (LDA-APF), GoP, and Linear Discriminant Analysis–Mel-frequency cepstral coefficients (LDA-MFCCs) for error detection were defined in [Strik et al. \(2009\)](#). The LDA classifier had a higher accuracy compared to the other classifiers in discriminating plosives from fricatives (the velar fricative /x/ and the velar plosive /k/).

[Ryu et al. \(2017\)](#) consider the best articulation features set from the well-known features in the previous studies to build their system on automatic pronunciation assessment. In the assessment modeling, the articulatory Goodness of Pronunciation (aGoP) was used for estimating the phoneme posterior probability.

[Tao et al. \(2016\)](#) investigated three ASR systems for nativeness evaluation in their study: a GMM-HMM system, a DNN-HMM system, and a GMM-HMM system using DNN for feature extraction. The feature sets were categorized into fluency, rhythm, pronunciation, grammar, and vocabulary for segmental (at the consonant and the vowel level) and suprasegmental measurements of non-native speech in automatic non-native speech assessment. We will base our work on these features in our ASR system.

2.1.1 Holistic pronunciation error detection

At the early stages of using ASR for pronunciation quality assessment, the preferred method was to consider the whole phrase without pointing to the error type and to perform the assessment with the help of a hidden Markov model ([Eskenazi, 2009](#)). Suprasegmental features such as pauses, intonation, stress, and speech rate were evaluated on proficiency in earlier studies. The earliest work on pronunciation evaluation to accurately predicting scores (similar to Oral Proficiency Interviews (OPI)) was performed by [Bernstein et al. \(1990\)](#). The system was used to assess non-native English proficiency of telephone data (conversational speech). The output scores from the system were compared with the scores of expert judgments of proficiency.

In [Minematsu \(2004\)](#), automatic scoring of individual learners’ pronunciation was investigated using the distorted phonemic method. The distortion metric method was used to measure the phonetic structures’ differences among speakers (native American English speakers and Japanese learners of English). This approach was different from the common features such as

confidence scores, Log-Likelihood Ratio (LLR), and log-posteriors applied to score speech in this field. The above studies are a small set of examples that are related to suprasegmental features. The holistic pronunciation error detection is not the focus of this study. Unlike these studies, we focus on pinpoint error detection, in which we observe the specific phoneme-level errors made by second language learners.

2.1.2 Pinpoint pronunciation error detection

Pinpoint error detection refers to identifying errors at the segmental levels (phone levels) (Chen and Li, 2016). Segmental features have been a subject of phonetics from the 1950s up to now. The most common segmental features for investigating the pronunciation of L2 learners on automatic speech assessment are **consonant features:** *stop closure duration*, *aspiration*, and **vowel features:** *vowel duration*.

The topic of interest in this thesis is the mispronunciation assessment at the phone level. Therefore, the representative research related to this topic will be discussed in this section. There are various ways to detect problematic phones in a speech at the phone level, which leads to an increase in precision of pronunciation assessment (e.g., Kim et al., 1997; Franco et al., 1999). There has been a large amount of research in pronunciation error detection (segmental level) since 1990. In one of the first studies on assessment of pronunciation at a phone level, Witt (1999) found that the scoring accuracy for the assessment of errors was 80-92% for non-native English speakers (73 speakers).

There are different approaches to the detection of mispronunciation at the phoneme level. Based on the previous studies (Kim et al., 1997), one of the first and simplest approaches is to segment the duration of phones for making the assessment. Phone duration is obtained from Viterbi alignments.

GoP, an approximation of the probability of the target phoneme, is a very similar approach to log-posterior probabilities and is one of the most common scoring approaches that is used to detecting mispronounced phonemes and measuring the goodness of learners' pronunciations (Witt and Young, 1997).

The system output is acceptability and unacceptability of the pronounced phoneme. To produce a GoP for non-native speech, earlier approaches (e.g., Kim et al., 1997) used a native model and this model has a higher correlation with human annotators in longer segments (such

as sentences and paragraphs) than shorter ones (such as phones).

Early research by [Franco et al. \(1999\)](#) used two kinds of acoustic models to conduct automatic mispronunciation detection: (i) a model trained on native-speaker pronunciation, and (ii) a model trained on non-native speech. They used an acoustic model to calculate the log probability for each predicted phone in both models. They then calculated the difference between these two probability scores and used it as a metric for rating the pronunciation's quality. The result showed that LLR had a better performance than the log-posteriors method. According to [Kim et al. \(1997\)](#), posterior probabilities and log-likelihood scores were the methods that were most correlated with word and phone level human assessments of pronunciation.

As an example of the earliest work in this field, we can refer to the FLUENCY system ([Eskenazi, 1999](#)) to detect pronunciation problems at both the phonetic and the prosodic level. They used the SPHINX-II speech recognizer to evaluate and detect phone errors and prosodic information (namely prosodic problems) of non-native speakers of French, German, Hebrew, Hindi, Italian, Mandarin, Portuguese, Russian, and Spanish, who learned English as their second language. It was reported that using ASR technology while learning a foreign language can reduce embarrassment and enhance learning for learners.

In this chapter, key ideas and existing systems were introduced. All in all, most of the research in this field uses ASR systems trained on native speech data to quantify the L2 learners' pronunciations errors in phoneme and word levels.

3 Description of data

The following sections describe the corpora and phone set that we used to identify pronunciation errors made by Italian learners of English, in addition to, some descriptive statistics of the number of phone errors in each corpus.

3.1 Speech corpora

The purpose of ASR is to recognize speech, and in order to reach high performance, the system needs hundreds of hours worth of recorded and annotated speech, coming from a large number of speakers with well-balanced genders, ages, and places of birth (Kato et al., 2020). In this study, we used two non-native English labeled corpora for developing algorithms capable of detecting mispronunciations and the description of each corpus will be given below; (1) a corpus of Italian children (ChildEn) (Batliner et al., 2005), (2) a corpus of Italian adults (ISLE) (Menzel et al., 2000), who are learning English as their second language. In this project, the detection performance of the ASR system was evaluated using the corpora, as mentioned earlier.

Corpora with proper labeling and agreed evaluation protocols are significant for the development of technology in the ASR field. Detecting mispronunciations in ASR requires corpora with labeling at the phonetic level: we selected these two corpora because they were manually labeled in terms of pronunciation quality by humans, and both were manually transcribed at phone-level, highlighting differences from the expected pronunciation.

3.1.1 ChildEn corpus

The ChildEn corpus¹ (Table 1) contains English sentences read by Italian children. The corpus consists of 5,268 utterances, with an overall duration of 3h:28m:26s from 78 children (44 males and 34 females) at about ten years of age. The selected students had been studying English at school for 3 or 4 years. The text was presented on a computer screen, and the child was asked to read the prompted text aloud. The recorded speech in the ChildEn corpus was divided into two sets: imitated and read speech.

In the imitated speech, the children (53 speakers: 26 boys and 27 girls) were provided with

¹This corpus was designed and collected by ITC-irst, <http://www.itc.it>

reference pronunciation from a native adult speaker and the children were asked to imitate the reference utterance, they had just listened to it. The duration of the sampled speech signals is 2h:8m:43s. The read speech (25 speakers: 18 boys and 7 girls) was recorded without any reference pronunciation. Each child was asked to read aloud a set of texts consisting of isolated words and short sentences. The duration of the sampled speech signals is 1h:19m:43s. English native annotators evaluated the pronunciation of words produced by Italian learners. The chosen phone set was the Speech Assessment Methods Phonetic Alphabet (SAMPA)¹. The frequency of pronunciation scoring is illustrated in Figure 3 in chapter 5.

3.1.2 ISLE corpus

ISLE (Table 1) is a non-native speech dataset which contains utterances recorded by intermediate-level adult German and Italian learners of English who read English sentences. The audio (WAV format) files are 17h:54m:44s in total. The Italian section contains 23 Italian speakers. The audio file has both sentence and annotation file at word and phoneme levels. Besides the word and phoneme annotations, the annotators scored the speakers on a 5-level scale based on speaker pronunciation proficiency. The data was transcribed using Entropic’s UK English phone set (Power et al., 1996).

Table 1: Information regarding the selected corpora

Information	ISLE	ChildEn
Language	German & Italian-L1 (learn English)	Italian-L1 (Italian, German & English)
Level	Intermediate or advanced learners	Elementary learners
Speaker	23 German & 23 Italian learners	78 Italian learners
Data	Read sentences	Read and imitated sentences
Hours	11,484 utterances, 17h:54m:44s	5,268 utterances, 3h:28m:26s
Age	Adult	Children
Annotation	Word & phone level	Word & phone level
Mispronunciation	Deletion/Insertion/Substitution	Deletion/Insertion/Substitution
Labeling	ASCII-based	SAMPA

Table 2 reports training and test data of Italian speakers in both ChildEn and ISLE data

¹<https://www.phon.ucl.ac.uk/home/sampa/>

sets. The data set is around 21 hours in total across both corpora and training/test sets. Furthermore, speakers between the training and test sets do not overlap.

Table 2: Statistics about the speech data sets used for ASR in this study

Corpus	Utterance	Duration
ChildEn-Train	4052	2.5h
ChildEn-Test	1215	1h
ISLE-Train	8239	14h
ISLE-Test	3245	3h

3.2 Choosing a phone set

According to [International Phonetic Association \(1999\)](#), phonemes are a set of vowels and consonants in a language or dialect. The set of phonemes are the smallest perceptible discrete unit of a speech sound that carries meaning in a language, and a phone is an instance of a phoneme, for example, an utterance can be described as a sequence of phones and silences.

Table 3 refers to information regarding the Italian and English phonemes. As observable in Table 3, Italian has 50 phonemes, of which 7 are basic vowels, and the rest are consonants. Table 15 in Appendix B shows the complete list of the phonemes used in our study. For each phoneme, we provided the IPA symbol. These phones contain both English and Italian phonemes.

In this study, we created a phone list of each phoneme that includes not only canonical pronunciation but also every possible mispronunciation of English phones by L2 learners (see Appendix B). In other words, we have created a phoneme dictionary that contains both English and Italian phones in order to be able to capture all possible mispronunciations. Through speech analysis, we want to detect and identify the words and phonemes spoken by converting the audio signal into an informative format. This, in turn, allows us to convert the sound of speech into meaningful information that can be used to evaluate the proficiency and fluency of the recorded audio.

Generally, learners of a foreign language with different phonological systems (such as different vowels and consonants) have problems producing the new language correctly, and making the appropriate distinction between their L1 and the L2. As was shown in Table 3, the Italian

Table 3: Language characteristics of English & Italian

Language	Language family	Morphological type	Word order	Number of Phonemes
English	Germanic Indo-European	Fusional	SVO	Vowels = 11 (Monophthongs), 8 (Diphthongs) Consonants = 24
Italian	Roman Indo-European	Fusional	SVO	Vowels = 7 (only Monophthongs) Consonants = 43

sound system has fewer vowels, and more consonants (such as geminate consonants¹: /gg/, /nn/) compared to the English sound system. Figure 2 shows the vowel chart of both English and Italian. The symbols in the charts are IPA symbols for each sound. The shape of the chart represents the side-view of the mouth. This chart shows two dimensions of movement of the tongue in the mouth: (i) Vowel Height: up/down movement, and (ii) Vowel Backness: forward/backward movement. According to Browning (2004), Italian and English share only 40% of their phones; in particular, the differences are the affricates, nasals, and liquids. Therefore, we expect to see more phonological interference from Italian when L2 learners need to pronounce the phones /dz/, /ng/, and /r/ in English. Moreover, in Italian, the relationship between spelling and pronunciation is straightforward compared to English. For example, in Italian the letter 'a' is pronounced /a/, but in English, pronunciation, and spelling are not strictly related to each other. For example the letter 'u' is pronounced /u/, /V/ or /ɜ/ in English².

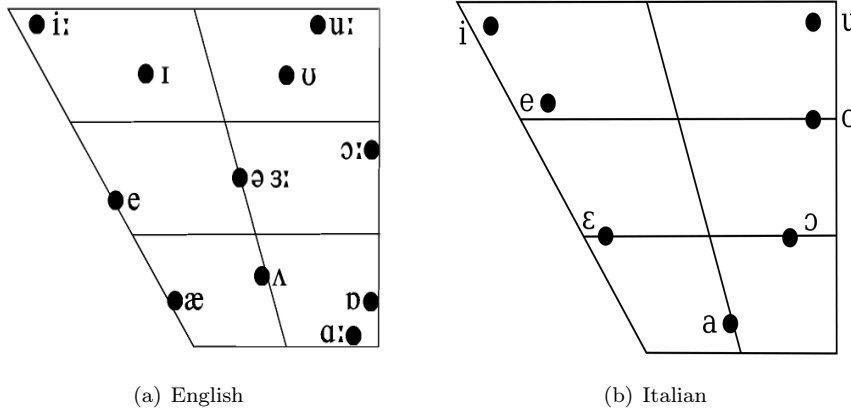


Figure 2: Vowel charts of English (left graph) & Italian (Right graph).

The Italian vowel system (only 7 monophthongs–Figure 2) is simple in contrast to the English vowel system (11 monophthongs and 8 diphthongs); this can cause mispronunciation for Italian

¹Geminate consonants are written with a double letter and refer to a longer pronunciation of the consonants in comparison to a single consonant.

²<http://archive.is/zsxA>

speakers who are learning English as their second language.

In the following Section 3.3, summarizes preliminary statistics extracted from both corpora and illustrates the output of the most common English pronunciation errors, such as frequency of errors per corpus, a list of most frequent errors along with phones per corpora with graphs and a detailed explanation for both child and adult speakers.

3.3 Phone errors: ISLE & ChildEn

Figure 3 shows the GoP of speech, which summarizes the judgments of annotators on speakers' proficiency level. Moreover, this output will be used to monitor the error process during the study. Most speakers show poor pronunciations (14636 phones labeled as "poor"). Table 4 reports the number of word and phone errors for each corpus of Italian speakers who learn English as their second language. We see a higher number of word and phone errors for ISLE because the number of hours of speech data is higher for ISLE in comparison to ChildEn. Furthermore, the list of common words (159 words) that occurred in both ISLE and ChildEn is summarized in Table 16 (see Appendix B).

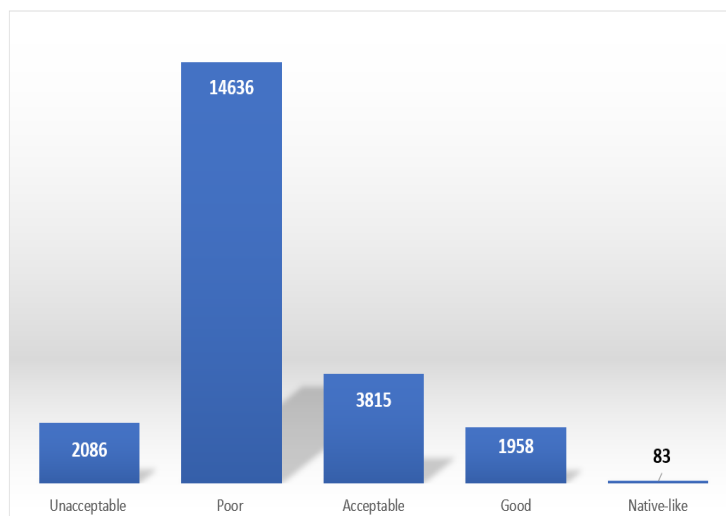


Figure 3: Frequency of degree of "goodness" of pronunciation

Table 4: Information of words & phone errors in ISLE & ChildEn

Folders	Word errors (Italian)	Phone errors	Hours	Utterance
ISLE	1006	6332	17h:54m:44s	11484

Table 4 continued from previous page

Folders	Word errors	Phone errors	Hours	Utterances
ChildEn	—*	3723	3h:28m:26s	5269

* There is no information regarding the errors at the word level in this corpus.

The next table (Table 5) shows the list of most frequent words in both corpora that lead to the most frequent phone errors for L2 learners. All pronunciation errors at words and phones levels were counted and sorted by frequency to determine which errors were the most frequent in each corpus.

Table 5: Most frequent words in ISLE & ChildEn

ISLE		ChildEn	
Word	Frequency	Word	Frequency
THE	6816	THE	349
A	5381	A	290
TO	3855	IS	251
OF	3090	I	152
I	2336	ARE	140
IN	2317	MY	117
AND	2044	GOT	105
NOT	1895	LIKE	100
IS	1652	YOUR	96
SAID	1578	IN	91
THIS	1239	WITH	75
FOR	957	SMALL	68
THAT	864	TELEVISION	66
LIKE	831	OF	62
METRES	824	I'M	61
EVEREST	778	GOOD	61
THEY	731	THERE	58
ON	708	THREE	57
HUNDRED	688	YELLOW	54
WE	685	BIRTHDAY	53

Table 5 shows the top 20 of the most frequent words in both ISLE and ChildEn, including content and function words. For example, the words “**there**” and “**like**” in ChildEn and ISLE were the most frequent word errors that happened in child and adult pronunciations based on the reports of both corpora. The word “**there**” contains the phoneme /dh/ for the letter “th” which is not available in the Italian phoneme set, and the word “**like**” was often pronounced with an extra phone /e/ on end. As mentioned earlier in chapter 4, the relationship between the spelling and pronunciation of words is straightforward in Italian.

3.3.1 Corpora statistics

The following figure (Figure 4) depicts the most frequent errors at the phone level based on the manually annotated dataset. Figure 4 tracks down the most common errors from ISLE and ChildEn at the phoneme level errors based on substitutions, deletions, and insertions. These statistics were created by calculating the errors between the gold standard annotation and the output from the acoustic model.

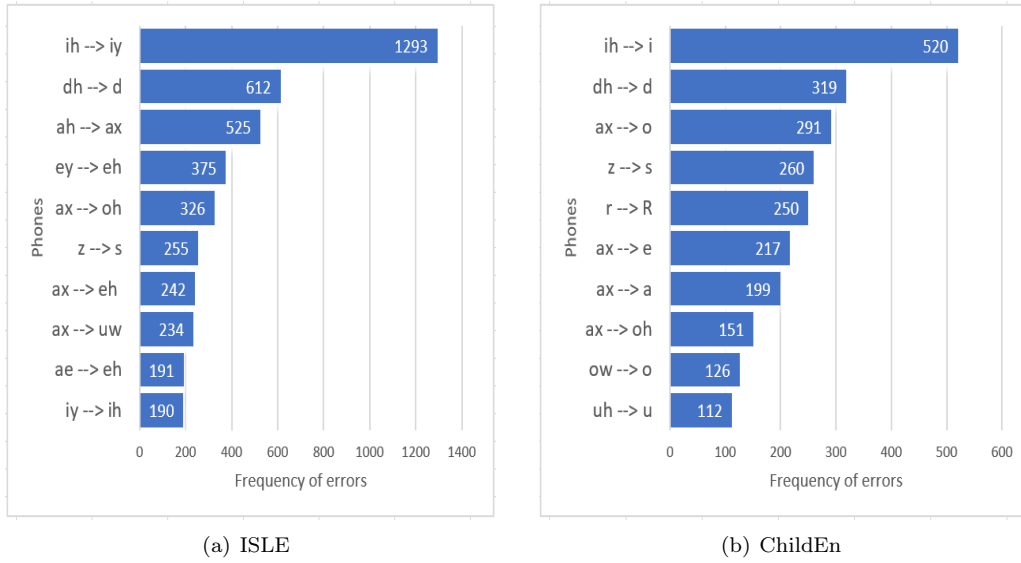


Figure 4: Phoneme errors distributions of ISLE (left graph) & ChildEn (Right graph).

Figures 9(a) and 9(b) (see Appendix B) describe the phoneme distribution in ISLE and ChildEn. According to Figure 4, the most common mispronunciations for Italian speakers occur when the English target phone is not in the phoneme set of Italian (e.g., English phoneme /ax/). These phones are not contrastive for Italian speakers, which leads to the speaker attempting to substitute the phonetically-closest phone from the Italian phoneme set or to delete that phone. By “close”, we mean that the acoustic signal has similar formants in both languages or the orthographic representation (spelling) is similar to spelling in Italian.

4 Material & Method

In the following sections, we describe the methods (language model and acoustic model) used to identify pronunciation errors made by Italian learners of English, to develop an ASR system to help in improving their L2 pronunciation.

4.1 Speech file transcriptions: Ideal, Manual, & ASR

For this study, the following three transcriptions at the phone level were considered: Ideal, Manual and ASR output (see Table 6).

- **Ideal** (reference) refers to the automatically time-aligned phonetic transcription of the sentence uttered, given the ideal transcription of each word.
- **Manual** (reference) output is related to a hand-corrected version of the ideal data at the phoneme level. Trained annotators corrected files manually by using Praat tool (Boersma and Weenink, 2016), noting substitution, insertion, and deletion at the phoneme level. The manual output contains the manually corrected time-aligned phonetic transcription and can be considered the best possible transcription at the phone level (Table 6).
- **ASR output** is the result of ASR processing given a set of phones as input. ASR processing is trained on audio corpora and some linguistic information at the phone level.

Given these three representations, the comparison between specific pairs of phone sequences can provide us the following information (see Table 22 or Table 23):

- **Ideal vs. Manual output:** this gives us a true map of the errors made by Italian speakers when trying to speak English;
- **Ideal vs. ASR output:** this gives us a feeling of what an automatic system can detect for a new utterance, where no manual annotation is available;
- **Manual vs. ASR output:** this tells us how well an automatic system detects errors.

For the preliminary part of this study, the error frequency of phones was evaluated by comparing it to ASR (**test/hypothesis**) output. Therefore, the comparison will be between reference phonemes with hypothesis phonemes, and based on our purpose; the comparisons are

as follows: MAN & ASR, IDE & ASR. The comparison of MAN & IDE added to the output for further evaluation. The phoneme level most common error estimation was conducted for each pair, and the result can be seen in Appendix C. According to previous studies in linguistics, pronunciation errors can happen both at segmental and suprasegmental levels. Errors that occur at the segmental (phonetic) level are divided into three categories: substitution, insertion, and deletion (Peabody, 2011). The phoneme level matrix was computed as follows:

- Obtained the canonical (reference) phoneme level transcriptions of the speech data
- Obtained the phoneme level ASR output transcripts as a hypothesis (test) of the speech data, and the ASR output transcriptions were aligned with the canonical/reference transcriptions (here Manual and Ideal).
- Furthermore, the probability was computed per each phone.

As mentioned in Chapter 2, the mispronunciation was categorized into three kinds at the phone level: substitutions, insertions, and deletions:

- Substitution: a phoneme is replaced with another.
- Insertion: an extra phoneme is inserted.
- Deletion: a certain phoneme is deleted.

The following refers to the examples for each error type in the corpora. A sample of manual (MAN), ideal (IDE), and ASR annotations for the sentence “**I said white not bait**” are as follows:

Table 6: Sample transcription: Manually (MAN), Ideally (IDE), & ASR

MAN: sil ay s eh d w ay t n oh t b ey t sil
IDE: sil ay s ax d w ay t n oh t b ey – sil
ASR: sil ay s ax d w ay t n oh t b ey – sil
 Silences were marked as “sil”.

The vowel /ax/ was substituted by the accurate vowel /eh/, and the consonant /t/ was deleted in ideal and ASR outputs. It means that the speaker mispronounced that vowel and replaced it with another vowel, close to the accurate vowel. The phone /t/ was not pronounced by the speaker, which refers to the deletion in this sample.

4.1.1 Sample

Generally, the sound wave representation of spoken words has two axes: time on the x-axis and amplitude on the y-axis. Figure 5 illustrates a sample sentence pronounced by a child, showing the speech waveform (top), the spectrogram (middle), and three text tiers (bottom) that report different segmental information in terms of words, phones, and some other information, along with their time boundaries. The spectrogram is a graphical representation which has three axes: time, frequency, and amplitude, where in 2-dimensional graphs, the amplitude is approximately visualized using a darker shading. Both corpora (ISLE and ChildEn) provided orthographic transcriptions at the word level for the speech input.

Figure 5 refers to the annotation and pronunciation of the phrase “**A birthday cake**” by an Italian male speaker. Three tiers were defined for each sound file; tier1) word, tier2) phone, and tier3) word score. Generally, three types of annotations were added to each audio for both corpora: sentence-level annotations (e.g., score and intonation), word-level annotations (e.g., pause), and phoneme-level annotations.

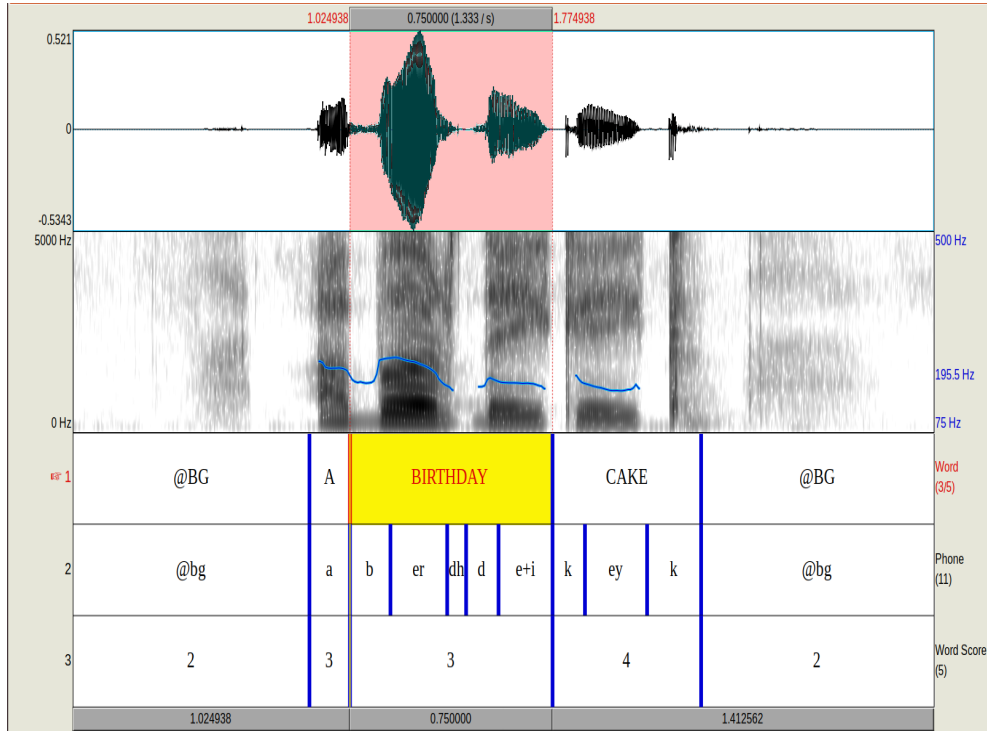


Figure 5: A TextGrid of manual annotation example from ChildEn corpus. **Top:** Raw waveforms of (x-axis: time; y-axis: amplitude). **Middle:** Spectrograms (x-axis: time; y-axis: frequency; shading: amplitude (energy), darker means higher). **Bottom:** Word-level annotation of the signal.

4.2 Basic framework of speech analysis

In this section, we will give a brief outline of the basic framework of speech recognition and how it works. A simplified diagram of an ASR system is shown in Figure 6. The input is a waveform and will be analyzed within short frames (20 ms). These frames are based on spectral information that is supposed to be stationary within each frame.

Frame-based analysis of the speech signal is done frame by frame, and the 1-dimensional time frame signal is converted to a 2-dimensional time-frequency image, called spectrogram (see Figure 5). Acoustic features can be extracted from this spectral information that can be used to identify phonemes in speech. In the next step, feature extraction of input will be sent to a recognizer with three main components (language model, pronunciation model, and an acoustic model) (Rabiner, 1989). Both acoustic and language models are statistical, i.e., they are trained on data (speech and text, respectively).

Typically, acoustic and language models need to be merged in a search space that the decoding algorithm then explores in order to find out the best path that corresponds to the word sequence (i.e., the sequence that maximizes some joint (acoustic and linguistic) probability). Acoustic models have the task to identify the phone(s) that best match the given sounds of the speech input speech, while language models assign a probability to each possible word (or phone) sequence.

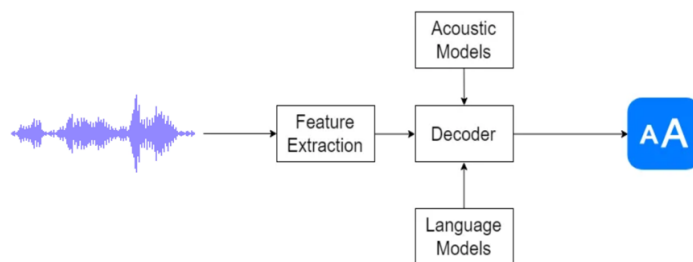


Figure 6: Components and models for ASR

The acoustic model is used to classify the acoustic frames into phones; additionally there is a language model (an n-gram model) trained on a big text corpus. There is also a grapheme-to-phoneme model that describes how words can be converted into phones. All of these components are combined into a Weighted Finite-State Transducers (WFST), which is used with the acoustic model's predictions to create the best possible word-level transcription for the sound. The

acoustic model consists of Gaussian mixture models (GMMs) or deep neural networks (DNNs). The language model and the acoustic model are then used to identify the word or phone sequences of a given audio segment. The language model converts phoneme sequences from the acoustic model to word sequences.

An ASR system converts an acoustic signal into a sequence of words based on some models, patterns, or algorithms (Errattahi et al., 2018); in other words, it automatically converts from speech to text (Kato et al., 2020). ASR technology plays an important role in many language learning tools and CALL systems for evaluating and improving pronunciation for L2 learners (Neri et al., 2003; Witt, 2012). The most common evaluation metric of ASR systems is Word Error Rate (WER), which calculates the proportion of word errors to words. One of the first topics for evaluating acoustic models, the phoneme-level error rate can be used. The phoneme-error rate calculated the log-likelihood of a predicted phoneme given the acoustic signal (e.g., Kim et al., 1997; Kawai and Hirose, 1998). However, for this task, a more useful metric is the Phone Error Rate (PER); this metric calculates the number of phones mispronounced, with respect to the total number of reference phones.

For the ASR system, we used Kaldi¹, the open-source speech recognition toolkit, to generate time-aligned versions of the audio file from orthographic transcriptions. As mentioned in the related work (Chapter 2), we use WER to evaluate the performance of the ASR system. We compare the transcript produced by the system against the target transcript. We also report results using PER.

4.2.1 ASR evaluation

Our ASR system’s goal is to evaluate the error rate and provide criteria based on different systems to measure the performance of learners according to their error rates. In other words, our ASR system has two aspects: scoring and feedback.

Scoring: Language teachers are the human raters for learner’s pronunciation. However, with ASR’s help, the rating will be more consistent and less time-consuming for evaluating non-native speech. In the ASR system, the first step is to identify the beginning and end of the speech (input) and phonetic segmentation. Pronunciation scores are computed from learner’s speech by comparing it to native-speaker speech. The required databases for pronunciation

¹Kaldi: <http://kaldi-asr.org/doc/>

scoring are as follows: native-speaker corpus as a pronunciation reference, non-native corpus to training the system, and human ratings for evaluating pronunciation skills.

Feedback: According to [Xiao \(2018\)](#), feedback is divided into scoring feedback and visual feedback. As stated in previous research (eg., [Menzel et al., 2000](#); [Neri et al., 2002](#)), the combination of audio feedback and visual feedback can improve the pronunciation of less-proficient learners; however, the auditory feedback is more effective for advanced learners. With the help of visual feedback, learners can modify their speech. The visual feedback can be used to evaluate suprasegmental aspects by using arrows that show learners which part of speech needs to be raised or lowered.

For assessing the performance of a phoneme, the phone error rate (PER) will be used. The PER takes into account the errors related to phoneme substitutions (S), phoneme deletions (D), phoneme insertions (I), and P stands for the number of phones. Generally, PER sums the three types of error. If there are P phones in the reference transcript, and the ASR output has S, D and I, then multiply by 100:

$$PER = \frac{S + D + I}{P} * 100 \quad (1)$$

$$Accuracy = 100 - PER\% \quad (2)$$

Eq. 2 shows how accurate a speech recognizer is.

[Witt \(2012\)](#) provided a list of pronunciation errors that can be considered in ASR studies. Phonemic deletion, phonemic insertion, and phonemic substitution are the pronunciation features used in this project.

4.2.2 Language model: N-Grams

The basic idea of language models is to provide a probability of a sentence or sequence of words and predict the upcoming words. In speech recognition tasks, the probabilities calculation is important in which these probabilities are used to identify words in unclear and noisy input. For a speech recognizer, it is essential to have the probable sequence of words or phones to realize the sentence that you said was *I said* “**white**” not “**bait**”. Based on the estimation of probabilities, the recognizer decides the word “**white**” is more probable than the word “**bait**”. The simplest probabilistic model is the n-gram (e.g., unigram, bigram, trigram, 4-gram, etc.):

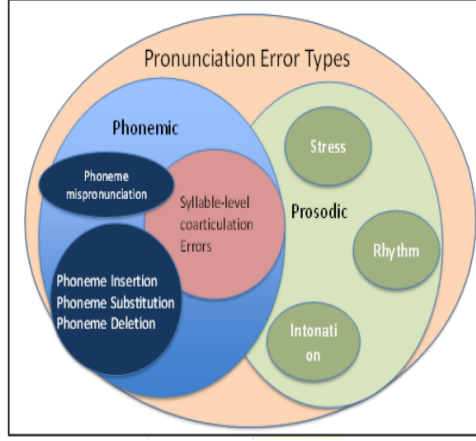


Figure 7: Pronunciation error types by categories (Witt, 2012).

it is based on sequences of words or phones (Jurafsky and Martin, 2018). This model was introduced by Markov (1913) in the form of Markov chains (bigrams and trigrams). For the output of this model, check Chapter 5¹.

$$P(\text{word}|\text{fixed window of history of previous words})$$

bigram model:

$$P(w_n|w_1^{n-1}) \quad (3)$$

This project uses n-gram stochastic language models, with n ranging from 1 to 5, and considering phones as units. 5-gram phone units, trained on the phonetic transcriptions of the training data led to the best results. Thus, no lexical information was used in this model.

The summary of the 20-top most frequent errors by using a 5-gram model are reported in Appendix C. Apart from the n-gram model, the novelty of this thesis is the application of a new language model (see next section 4.2.3) based on the most frequent errors (hereafter error model).

4.2.3 Language model: Error model

For the error language model, we used lexical information by providing the ASR with the canonical pronunciations of the (known) words to be recognized and other pronunciations obtained

¹We used an n-gram language model for the preliminary result based on the training data.

by applying phonetic rules. These rules were defined manually by looking at the most common errors resulting from the 5-gram phone model. Our error model is used to distinguish between critical (the most common) and non-critical errors as defined by the error rules for our ASR system. The second ASR system is based on the adapted pronunciations of the current lexicon. To train the error model, each transcription is aligned with all the answers in the manual transcription associated with the corresponding audio. We created a new lexicon with all the possible pronunciations of words based on the most common errors. The defined rules were summarized in Table 7.

Table 7: Error rules: Substitutions, Deletions, & Insertions

Substitution rules		Deletion rules		Insertion rules	
Original	Sub	Original	Del	Original	Insert
ih ->	ih/i/iy	n d ->	n d/	er ->	er r/R/
dh ->	dh/d	l d ->	l d/	aa ->	aa r/R/
ax ->	ax/a/o/oh	th d ->	th d/	ao ->	ao r/R/
z ->	z/s	s t ->	s t/	p l ->	p ax/o l
r ->	r/R	n t ->	n t/	b l ->	b ax/o l
uh ->	uh/u	ay k ->	ay k/	k l ->	k ax l
		ae k ->	ae k	g l ->	g ax l

As an example, the rule “**er** → **er r/R/**” means that our adapted lexicon ¹ will have the following transcriptions for the word “**HER**”:

- hh er (ideal phonetic transcription)
- hh er **R**
- hh er **r**

Table 8 shows the two examples from the adapted lexicon. The pronunciation column (**phonetic transcription**) refers to the actual pronunciation of the words, and the third column is about adapted pronunciations (**phonetic transcription**) based on the defined rules.

¹Table 24 shows the adapted lexicon of both ISLE and ChildEn. The new adapted lexicon has 382456 transcriptions and the original lexicon has 56101 transcriptions. This adapted lexicon extracted by applying the defined rule in section 4.2.3.

Table 8: Sample of adapted pronunciations for the words “Autumn” & “Happy” from ChildEn corpus.

Lexicon	Pronunciation	Adapted pronunciation
Autumn	ao t ax m	ao t a m ao t ax m ao t o m ao t oh m hh ao t a m hh ao t ax m hh ao t o m hh ao t oh m
Happy	hh ae p ih	ae p i ae p ih ae p iy hh ae p i hh ae p ih hh ae p iy

4.2.4 Acoustic model

The acoustic model is learned from a set of audio recordings and their corresponding transcripts. The ASR system is provided with the phonetic transcript ahead of time, and asked to align the acoustic signal to the phones in the transcript. For each frame of audio, the ASR system makes a prediction on what phone it believes that the frame represents. It also assigns a log-likelihood score to that prediction. For segmental error evaluation, forced alignment (the automatic process of determining the duration and time of occurrence for each word and phoneme (Ai, 2018)) with the acoustic model takes the most frequent errors of L2 speech errors (Figure 9). Table 9 shows the example of alignment for the sample sentence **A birthday cake**. The columns in Figure 9 is structured as follows:

- The first column is the speaker id.
- The second column and third show the start and end time (duration) of the each phoneme
- The fourth column refers to the representation of phonemes
- The last column contains the confidence scores output by the acoustic model¹.

¹The confidence metric was firstly introduced by (Franco et al., 1997) for the posterior probability output.

The Training happened on native and non-native speech data. The core of the ASR method is to use deep neural networks as an acoustic model for phoneme recognition and detection of mispronunciations. We used DNNs to extract information from the acoustic parameters of the speech signal for ASR. For the acoustic model, we trained a DNN on English and Italian speech data from the child and adult Italian speakers (ChildEn and ISLE corpora). Based on the Kaldi recipe, the selected features for our models for the ASR model are: (i) a GMM triphone model: Using MFCC acoustic features applying LDA, speaker adaptive training (SAT) (Miao et al., 2015), and (ii) i-vectors (Madikeri et al., 2016) of size 100: are stacked to 40 MFCCs.

Table 9: Example of forced alignment result from Kaldi toolkit: **A birthday cake**

Speaker_ID	0	0.08	sil*	0.83
	0.08	0.11	a	1
	0.22	0.08	b	1
	0.3	0.18	er	1
	0.48	0.03	th	1
	0.51	0.09	d	1
	0.6	0.15	ey	1
	0.76	0.17	k	1
	0.93	0.21	ey	1
	1.14	0.23	k	1
	1.37	0.19	sil	0.76

*Silences beyond sentence boundary and pauses are marked as “sil”.

5 Results

The results chapter of this thesis is organized into five sections:

- Section 5.1 details the implementation of our models for accurately detecting pronunciation errors at the phone level.
- Section 5.2 presents our ASR output and the accuracy of our language models.
- Section 5.3 closes with a summary of the main results of our models.

5.1 Detection output

In this study, we are concerned with identifying mispronounced phonemes by L2 learners and improving our ASR system. Moreover, we are also interested in identifying every single mispronunciation. Therefore, we did not consider GoP for our study. Consequently, we are interested in the PER and accuracy of our system.

For phoneme error detection, we used Kaldi to perform speech recognition using the dictionary of words and phonemes. The acoustic model was trained with mixed speech data (i.e., children and adults), and the utterances were force-aligned based on the newly adapted word transcription. In our ASR system, we considered the following procedures in the acoustic model; (i) speech recognition based on the phone-level n-gram model., (ii) forced-alignment based on the existing word transcriptions, including the transcriptions modified using the mispronunciation data., and (iii) Moreover, GMM classifier on using the native and the non-native acoustic model were used. In both cases, the ASR output is a sequence of phonemes, as explained in Chapter 4. The following sections are the result of comparing the output of phoneme sequences to the gold standard, which leads us to identify pronunciation errors.

5.1.1 N-Grams output

In this section, the output and PER for each n-gram model will be summarized. The ASR was run on 1-gram, 2-gram, 3-gram, 4-gram, and 5-gram (see Appendix C for the complete table). We first trained the algorithm introduced in Chapter 4 (Section 4.2.2) on our data to obtain the baseline values for the top frequent errors to determine the rules for the error model. The results of PER obtained with the 5-gram method are summarized in Appendix C.

We trained all the n-gram models, but we saw that performance was much better with the 5-gram language model. For the rest of the n-grams, we only report the PERs for ASR & MAN, ASR & IDE, and IDE & MAN in (see Table 10). The reason is that the 5-gram model captures more context than other models, and thus we chose to report in full detail only the 5-gram results based on the number of substitutions, insertions, and deletions (Table 11). For this reason, the 5-gram model was used for finding modified transcriptions. The PERs of 5-gram in ISLE and ChildEn are 42% and 38% for ASR & MAN comparison, respectively. According to Jurafsky and Martin (2018), the common n-gram models, when there is sufficient training data, are trigram, 4-gram, or even 5-gram. As we have enough data for training, we consider the output of 5-gram models to develop our error model rules. According to Table 10, the performance of the ASR systems that used n-gram models was low. In other words, the PER for each n-gram model was around 33-52% (see Appendix C).

Table 10: PER output of n-gram models

Corpora		PER				
		1-gram	2-gram	3-gram	4-gram	5-gram
ISLE	ASR & MAN	52%	48%	46%	44%	42%
	ASR & IDE	51%	47%	45%	42%	38%
	IDE & MAN	12%				
ChildEn	ASR & MAN	47%	49%	45%	40%	38%
	ASR & IDE	45%	47%	42%	36%	33%
	IDE & MAN	15%				

Table 10 summarizes the results from the n-gram models. The 5-gram model has better PER compared to other n-gram models: the range of PER for the 5-gram model is between 33% to 42%. Moreover, Table 22 and 23 (Appendix D) report the top most frequent errors that non-native English learners make and show where the ASR fails to recognize the phone correctly in both corpora, due to the student’s mispronunciation. Our ASR system was unable to recognize the top frequent errors reported in the following table. The model we were provided with was not optimized on this task, and training another model was outside of the scope of this work.

As previous traditional research in this field shows (e.g., Kruk, 2012), the majority of pro-

nunciation mistakes come from substitutions and deletions. Due to this fact, different implementations were considered to detect these errors automatically.

Based on the lower PER (see Table 10), we chose the 5-gram output to define the error rules for our error language model. The number of substitutions, insertions, and deletions for each output (ASR & MAN, ASR & IDE, and IDE & MAN) (see Chapter 4) is shown in detail in Table 11.

Table 11: Output of number of errors & PER based on 5-gram model: Total (T), Substitutions (S), Insertions (I), & Deletions (D)

PER	ISLE				ChildEn			
	T	S	I	D	T	S	I	D
ASR & MAN	50584	17912	11049	21623	21809	9840	2181	9788
	PER= 42%				PER = 38%			
ASR & IDE	45355	14310	11311	19734	17990	7541	2596	7853
	PER= 38%				PER= 33%			
IDE & MAN	14570	9561	1429	3580	8463	5377	368	2718
	PER= 12%				PER= 15%			

Based on the ASR output (Table 11) Italian children make more errors than adult learners of English. Their errors often come from consonant substitutions, while the mispronunciations of adult speakers come from the vowel system.

Although we used the n-gram model’s output to define our set of error rules, the performance of the n-gram model was reduced (52% PER). Therefore, we implemented a new language model based on the error rules extracted from the 5-gram model and then trained a new system using the adapted lexicon. The results of this new system are summarized in the next section.

5.1.2 Error model output

Thus far, we have tried to improve the PER to estimate the performance of the ASR system for mispronunciation recognition of L2 learners of English. The second model used the predefined rules (Table 7) from the 5-gram model to train native and non-native speech data to check if the output of our ASR system will be improved in comparison to the n-gram model. In other words, we applied the force-alignment by using the adapted lexicon.

The results of using the data on native and non-native acoustic models were summarized in Table 12. Interestingly, the new ASR system performs better in terms of PER for ChildEn. In other words, the error rules showed better improvement for the ASR system in using ChildEn. The reason for the low improvement in ISLE might be due to more complex and long sentences used in this corpus that lead to error propagation differently. The other possible reason might be due to the quality of the recordings (i.e., noise in the background). More details on these results can be found in Table 12.

Table 12: Output of PER based on error model

PER Output	ISLE		ChildEn	
	native	non-native	native	non-native
ASR & MAN	44%	43%	28%	28%
ASR & IDE	38%	38%	24%	23%
IDE & MAN	12%		15%	

However, based on the output summarized in Table 12, we can see no difference between the two models (native and non-native models) for both ISLE and ChildEn. Having quite close PERs for both native and non-native data illustrates the robustness of our ASR system, which fulfills one of the research goals of this project.

5.2 Evaluation

The purpose of this study is to find an ASR system with high accuracy for detecting mispronounced phonemes. Table 13 shows the statistics of error detection results comparing the error detected by our ASR system and errors reported by annotators for each corpus. The comparison of the performance of our models (n-gram model and error model) was done by calculating PER. We choose the best model based on the PER metric. The speech recognition phone error rates are typically greater for adults than children. Overall, the highest Average Accuracy (see Eq. 2 in Chapter 4) to date was obtained using the error model. The accuracy results range from 72%-76% for correct error detection of phones in ChildEn.

Table 13: Detection Accuracy: n-gram & error model

Corpora		5-gram	error model
ISLE	ASR & MAN	58%	56%
	ASR & IDE	62%	60%
	IDE & MAN	88%	
ChildEn	ASR & MAN	62%	72%
	ASR & IDE	67%	76%
	IDE & MAN	85%	

As mentioned in section 4.2.3, we decided to show only the detection accuracy of our native model of both n-gram and error models in Figure 8. Among these models, the system detects performs better in the error model for ChildEn. The PER difference between ISLE and ChildEn is mainly due to the acoustic model. It should be noted that we did not evaluate the differences between native and non-native models in terms of likelihood ratios; however, this is an avenue for future research.

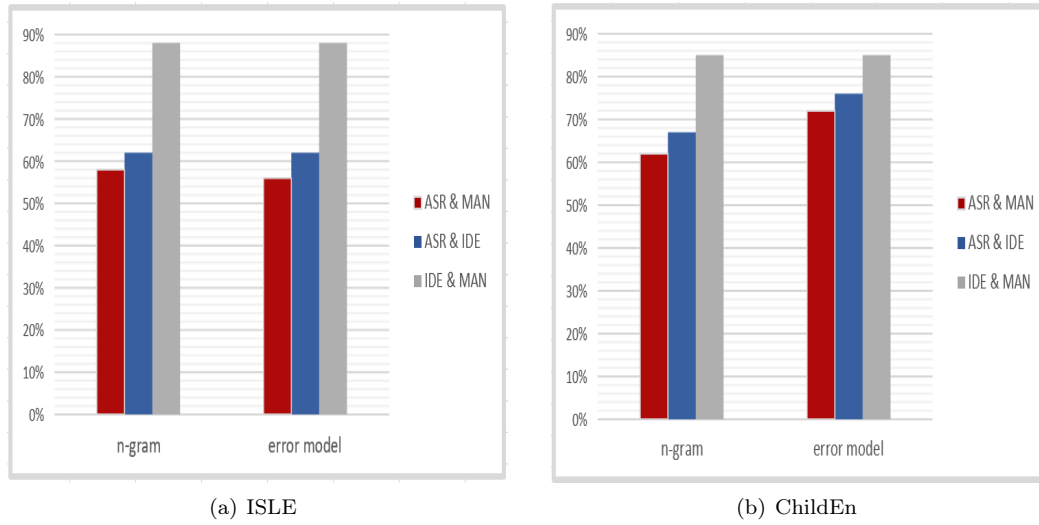


Figure 8: Overall detection accuracy for ISLE (left graph) & ChildEn (Right graph).

As mentioned in Chapter 4, the output of ASR & MAN helps us to evaluate how well our ASR system was able to detect errors in our data. Furthermore, the comparison between ASR & IDE output tells us about what the ASR system can detect from a new utterance without having manually annotated data. The accuracy of comparison of ASR & MAN output for

ChildEn is 72%, and ASR & IDE is 76%.

5.3 Summary

Our results show that using an error model can improve the performance of an ASR system. Our ASR models based on n-grams were not accurate enough due to their high PER (see Table 10). Given the output of our system described in the previous sections, the error model is effective and could improve the detection of mispronounced phonemes with high accuracy. However, we did not see a huge difference in the PER of the n-gram and the error models for ISLE corpus. This can be due to the poor quality of the acoustic model. However, the performance of our system on ISLE was poor, since we created the rules based on ISLE as well as ChildEn. The PER output of our native and non-native data were quite similar, shows that our system is sufficiently robust, and this was one of the goals of our study. In summary, the new ASR system with the error model showed much improvement in recognition accuracy of mispronounced phones with a relatively low PER, particularly for ChildEn. More details about the PER outputs can be found in section 5.2.

6 Final discussion & Conclusion

Based on earlier research in assessing pronunciation, the most reliable estimates came from the suprasegmental level, in which the assessment was based on paragraphs composed of several sentences for speakers’ overall pronunciation proficiency, instead of on smaller units. Focusing on the smaller units would allow students to focus more on specific aspects of their pronunciation. However, the evaluation of smaller units for pronunciation assessment has a higher uncertainty compared to the evaluation of longer units (Witt and Young, 2000).

As mentioned in the introduction chapter 1, for pronunciation training, especially for L2 learners, automatic pronunciation evaluation can play an important role, for example, by using ASR systems to evaluate the pronunciation of non-native input and quantify how close the speech is to a native-like pronunciation.

In this study, we worked on the language model - at the phone level - to extract the mispronounced phones and trained an acoustic model of native and non-native training examples for the detection of mispronunciations. One of the challenges in using non-native speech for automatic recognition is the diversity of allophones, accents, invented words by L2 learners, and longer hesitations. In this project, we identified high priority phones (i.e., the phones used for creating the error rules (see Table 7) and evaluated our system based on them. Since achieving a perfect native-like pronunciation is an unrealistic goal for adult learners, we focused on the specific mispronounced phones summarized in Table 7 (Chapter 4).

In this study, we compared the PER of our ASR system using native and non-native speech data in order to determine the validity of our error model hypothesis. The innovation of our system is to use error rules in our language model (error language model) based on the most common errors that were seen in L2 learners’ speech, in order to improve our detection system.

Previous works of WER/PER in speech recognition report that WER for human annotators on native data is around 5% in comparison to WER on non-native speech data, which is as high as 30% for automatic annotations (Zechner et al., 2009). Our ASR system may not be able to perform better than this human-level performance from human annotators in detecting mispronounced phonemes in non-speech data. The best recent report of performance on WER comes from an ETS project, in which the authors trained an ASR system on 800 hours of speech data. The reported WER was 28.5% (Chen et al., 2018).

In fact, the outputs of our ASR systems showed that by applying n-gram and error models, the model proposed in this work obtained the lowest PER and better accuracy for both ISLE and ChildEn. In other words, in our system, the model that produced the best PER is the error language model with a PER of 23%, which is better than the rate in previous studies.

This model was trained on the phone errors observed in our corpora. We created a set of rules to take into account these errors in our corpora and implemented our system based on the adapted lexicon. From the output of our ASR system, it is clear that the selected architecture is more capable of detecting mispronounced phones.

Table 14 shows examples of reference transcriptions (i.e., IDE & MAN) with their ASR output based on the error language model. We can see that the ASR system, by considering the error language model, performs better at recognizing the phones. We can see that our ASR system selected the transcription defined in the new adapted lexicon based on our error model (last column of the table). For example, the English sound /r/ was not selected by our system, and instead of it, the Italian sound /R/ was replaced for the word “**free**” that is the same with the manually annotated file (MAN). For the word ‘**station**’, the ASR output (n-gram model) was not able to recognize the sound /ax/ although our ASR output based on the error model identified it as the Italian phoneme /a/.

Table 14: Examples of English reference transcriptions (IDE & MAN) & the selected transcription of our ASR system

Word	IDE	MAN	ASR (n-gram)	ASR (error model)
JACKET	/jh ae k ih t/	/jh ae k ax t/	/jh ae k ih t/	/jh ae i t/
FREE	/f r iy/	/f R iy/	/t r iy/	/f R iy/
STATION	/s t ey sh ax n/	/s t ey sh ax n/	/s t ey sh n/	/s t ey sh a n/

Improving ASR will provide more speaking possibilities for learners to learn a new language. Using ASR can improve the traditional class to a more learner-centered environment with less anxiety. The part that still needs more attention is the kind of feedback that an ASR system wants to generate and to compare a teacher’s feedback in terms of effectiveness and comprehensibility. Apart from the segmental aspect, the suprasegmental aspect also plays a role in pronunciation training, but the ASR for the suprasegmental part is less reliable and still

needs more work to improve it. The error detection system records the learners' utterances, and the phone-level transcription will be provided by the ASR-based detector. As the final step, a list of possible feedbacks is given to L2 learners based on their pronunciation errors. Our ASR system needs to be adapted to the mother tongue of the L2 learners since different L1s can cause different pronunciation errors. Our error rules were defined based on the common errors of Italian speakers who learn English as their second language.

6.1 Directions for future research

Finding errors is not the final destination. The next important step is to provide initiative feedback to learners which helps learners to correct their errors recognized by the ASR system. There are a number of requirements for providing constructive feedback: the feedback should be correct, immediate, and be in a useful form for learners by suggesting ways to address errors (e.g., [Eskenazi, 1999](#); [Neri et al., 2002, 2003](#)).

6.1.1 Feedback

Such as in all research, there are some parts of our work that could be improved or expanded on in future work. This last section addresses the potential direction for future research. Effective feedback for learners can be:

- Providing information for word and phoneme problem information.
- Providing pronunciation scores not only for each section and word but also for each phoneme.

Feedback on pronunciation is of key importance for learning and improving second language proficiency. It helps non-native speakers to identify their errors and to focus on improving their pronunciation performance. Ideally, feedback can be presented to learners by synthesizing the learner's own speech data in a combination of two ways: visually and perceptually. In addition, giving practical advice for individual phones can help the learner, such as, for example, "make the duration of the consonant shorter" or "try to bring the position of your tongue a bit forward".

6.1.2 Duration

Another potential work would be to regard speech rate of speakers and the duration of geminates (Italian) in the error language model, with the aim of improving the proposed approaches and their related results to provide a feedback for language learners. Duration of phones was not the scope of this study, although in previous research (e.g., [Truong et al., 2005](#)) the addition of duration slightly improve the accuracy of their models. We estimate that adding the duration into our model will change the accuracy, due to the fact that Italian geminates play an important role in distinguishing words in this language.

Apart from the aforementioned future directions, prosodic features (e.g., stress and intonation) can also be integrated with segmental level information to identifying individual errors. All in all, this project can be considered as the baseline research on the detection of mispronunciation errors of a second language learners that can help us to the improvement of the computer assisted language learning tools for learners.

Appendices

Appendix A List of Abbreviations

aGoP	articulatory Goodness of Pronunciation
ASR	Automatic speech recognition
CALL	Computer Assisted Language Learning
CAPT	Computer-assisted pronunciation training
DNN	Deep neural network
DNN-HMM	Hybrid Deep Neural Network-Hidden Markov Model
ETS	Educational Testing Service
GMM-HMM	Gaussian Mixture Model-Hidden Markov Model
GoP	Goodness of Pronunciation
L1	First Language (native language)
LDA	Linear Discriminant Analysis
L2	Second language
LM	Language model
LLR	Log-Likelihood Ratio
ISLE	Interactive Spoken Language Education
MFCC	Mel-Frequency Cepstrum Coefficients
OPI	Oral Proficiency Interviews
PER	Phone Error Rate
SAMPA	Speech Assessment Methods Phonetic Alphabet
WER	Word Error Rate
WFST	Weighted Finite-State Transducers

Appendix B Words & Phones list

Table 15: Phoneme list extracted from MAN with examples & their transcriptions

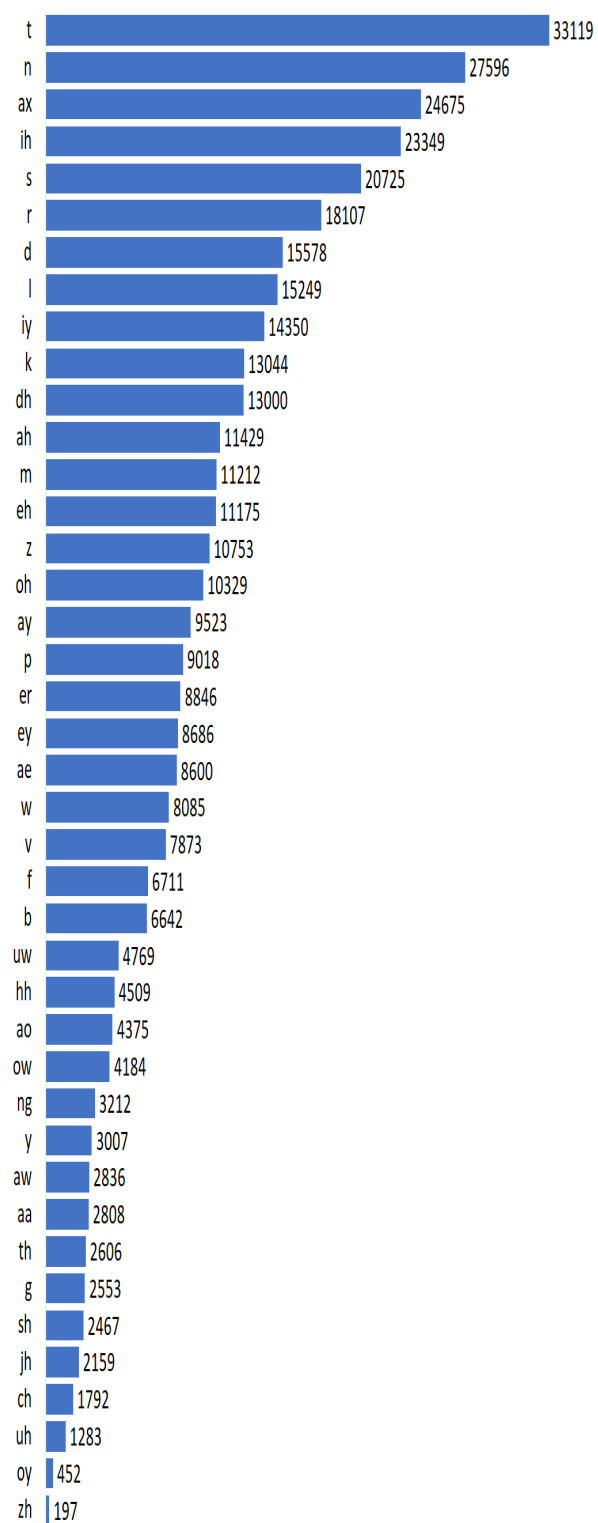
ISLE	ChildEn	Full list	Examples
–	a	a	casa
aa	aa	aa	car
ae	ae	ae	man
ah	ah	ah	cup
ao	ao	ao	water
aw	aw	aw	now
ax	ax	ax	father
ay	ay	ay	sky
b	b	b	bag
ch	ch	ch	match
d	d	d	dress
dh	dh	dh	with
–	dz	dz	arzilla
–	e	e	eroe
–	E	E	prendisole
–	ea	ea	there
eh	eh	eh	bed
er	er	er	girl
ey	ey	ey	day
f	f	f	family
g	g	g	big
hh	hh	hh	happy
–	i	i	iris
–	ia	ia	here
ih	ih	ih	kid
iy	iy	iy	green

Table 15 continued from previous page

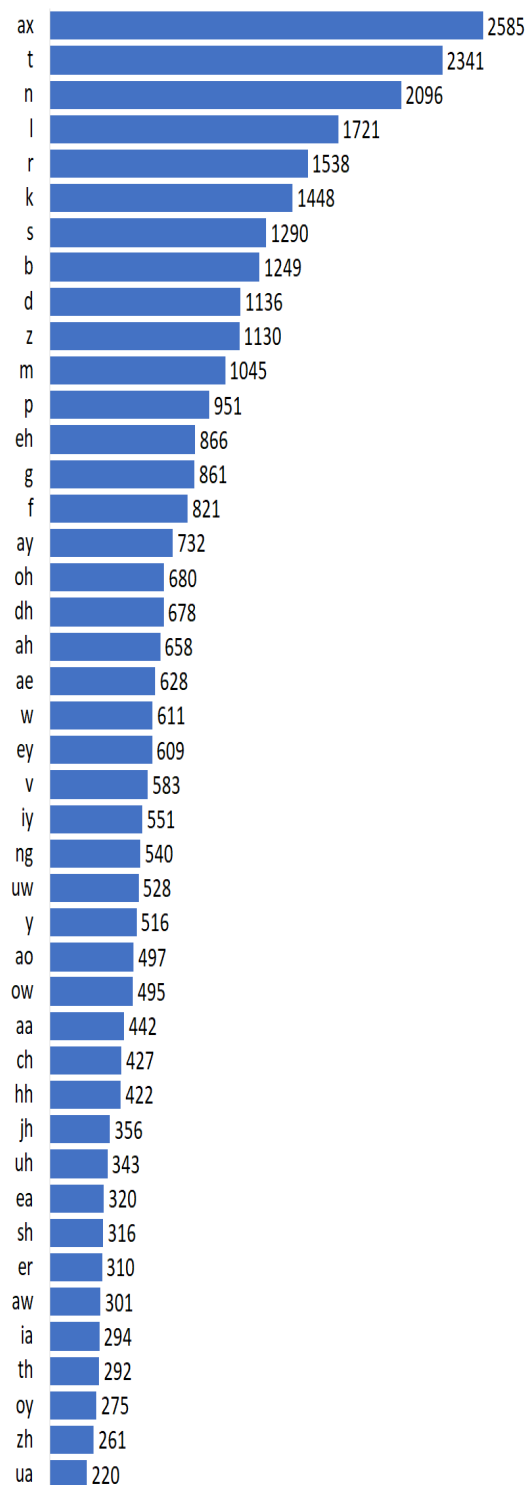
ISLE	ChildEn	Full list	Examples
–	j	j	ieri
jh	jh	jh	juice
k	k	k	school
l	l	l	like
m	m	m	arm
n	n	n	funny
ng	ng	ng	song
–	O	O	portaborse
–	o	o	non
oh	oh	oh	dog
ow	ow	ow	close
oy	oy	oy	oil
p	p	p	pet
–	R	R	perdono
r	r	r	road
s	s	s	sun
sh	sh	sh	nation
t	t	t	city
th	th	th	bath
–	ts	ts	zappa
–	tz	tz	puzzle
–	u	u	durata
–	ua	ua	europe
uh	uh	uh	book
uw	uw	uw	food
v	v	v	drive
w	w	w	wall
y	y	y	year
z	z	z	easy
zh	zh	zh	pleasure

Table 16: Common words in both ISLE & ChildEn

A	CHICKEN	FRUIT	JACKET	OUTSIDE	SITTING	TREE
ABOUT	CHILDREN	FULL	JUG	OVER	SLEEP	TWO
AM	CLOSE	GAMES	JUST	PARIS	SOME	VERY
AN	CLOTHES	GERMANY	KITCHEN	PARK	SPORT	VILLAGE
AND	COFFEE	GLASS	LAMP	PEN	STATION	VOICE
ANY	COMPUTER	GO	LATE	PEOPLE	SUMMER	VOICES
ARE	COWBOY	GOING	LIKE	PICTURE	SURE	WANT
AT	CUP	GOT	LONG	PIE	TABLE	WATER
BAD	DAY	GREEN	LOVE	POINTS	TAKE	WE
BAR	DO	HAS	MANY	POTATOES	TELEVISION	WEARING
BE	DON'T	HAT	MARCH	PUT	TEN	WE'RE
BEACH	DRINKING	HAVE	MATCH	RAILWAY	THAT	WHERE
BED	EASY	HE	MONTH	READING	THE	WHITE
BIG	EIGHT	HER	MOUSE	RED	THERE	WITH
BLUE	FARM	HE'S	MOUTH	RIVER	THERE'S	WONDERFUL
BOAT	FEW	HIS	MY	SAID	THESE	WORK
BOILED	FINGER	HOME	NEAR	SAY	THEY	YEAR
BOOK	FISH	I	NEW	SHE	THINK	YEARS
BOTTLE	FOOD	I'M	NEXT	SHEEP	THIS	YELLOW
BROWN	FOOT	IN	NOT	SHE'S	THREE	YORK
CAN	FOR	IS	OF	SHORT	TO	YOU
CAR	FRIEND	IT	ON	SHOULDER	TOO	YOUR
CHEESE	FROM	IT'S	OR	SINGING	TREE	



(a) ISLE



(b) ChildEn

Figure 9: Phoneme distributions of ISLE (left graph) & ChildEn (Right graph).

Appendix C PER: N-Gram output

Table 17: PER of 1-gram model

PER	ISLE				ChildEn			
	T	S	I	D	T	S	I	D
ASR & MAN	613144	21883	9665	29766	27029	11535	1837	13657
	PER= 52%				PER = 47%			
ASR & IDE	59133	21133	10025	27975	24574	10764	2170	11640
	PER= 51%				PER= 45%			
IDE & MAN	14570	9561	1429	3580	8463	5377	368	2718
	PER= 12%				PER= 15%			

Table 18: PER of 2-gram model

PER	ISLE				ChildEn			
	T	S	I	D	T	S	I	D
ASR & MAN	55163	21380	10138	25645	27988	11375	1868	14745
	PER= 48%				PER = 49%			
ASR & IDE	55037	20481	10600	23956	25549	10658	2182	12709
	PER= 47%				PER= 47%			
IDE & MAN	14570	9561	1429	3580	8463	5377	368	2718
	PER= 12%				PER= 15%			

Table 19: PER of 3-gram model

PER	ISLE				ChildEn			
	T	S	I	D	T	S	I	D
ASR & MAN	55261	20790	10592	23879	25596	10827	2014	12755
	PER= 46%				PER = 45%			
ASR & IDE	52911	19701	11037	22173	22927	9716	2410	10801
	PER= 45%				PER= 42%			
IDE & MAN	14570	9561	1429	3580	8463	5377	368	2718
	PER= 12%				PER= 15%			

Table 20: PER of 4-gram model

PER	ISLE				ChildEn			
	T	S	I	D	T	S	I	D
ASR & MAN	52752	19804	11174	21174	22899	10481	2448	9970
	PER= 44%				PER = 40%			
ASR & IDE	49569	17884	11618	20067	19678	8638	2934	8106
	PER= 42%				PER= 36%			
IDE & MAN	14570	9561	1429	3580	8463	5377	368	2718
	PER= 12%				PER= 15%			

Table 21: PER of 5-gram model

PER	ISLE				ChildEn			
	T	S	I	D	T	S	I	D
ASR & MAN	50584	17912	11049	21623	21809	9840	2181	9788
	PER= 42%				PER = 38%			
ASR & IDE	45355	14310	11311	19734	17990	7541	2596	7853
	PER= 38%				PER= 33%			
IDE & MAN	14570	9561	1429	3580	8463	5377	368	2718
	PER= 12%				PER= 15%			

Appendix D Sample output of 5-gram model: ISLE & ChildEn

The following tables 22 & 23 are the top frequent errors in our 5-gram model; Substitutions, Insertions & Deletions.

Table 22: Output of top frequent errors for ISLE corpus generated form 5-gram model

ASR & MAN			ASR & IDE			IDE & MAN		
Substitutions	Insertions	Deletions	Substitutions	Insertions	Deletions	Substituions	Insertions	Deletions
1162 *S_iy_ih	1937 *I_sil	12616 *D_sp	486 *S_er_ax	2001 *I_sil	12878 *D_sp	1312 *S_iy_ih	376 *I_t	2148 *D_ax
642 *S_ax_ah	791 *I_t	2557 *D_ax	423 *S_ax_ah	746 *I_t	1708 *D_r	641 *S_d_dh	219 *I_hh	379 *D_r
562 *S_eh_ax	707 *I_n	1965 *D_r	422 *S_ah_ax	710 *I_n	1211 *D_sil	522 *S_ax_ah	137 *I_d	261 *D_g
524 *S_oh_ax	531 *I_ax	1170 *D_sil	418 *S_ih_iy	577 *I_ax	868 *D_ax	375 *S_ey_eh	96 *I_k	176 *D_hh
482 *S_d_dh	496 *I_s	362 *D_d	416 *S_oh_ax	510 *I_d	349 *D_t	343 *S_eh_ey	77 *I_n	70 *D_b
463 *S_er_ax	482 *I_d	252 *D_g	384 *S_ey_ea	497 *I_s	226 *D_dh	327 *S_oh_ax	69 *I_y	57 *D_p
355 *S_sp_t	465 *I_ih	216 *D_t	370 *S_eh_ax	491 *I_ih	217 *D_d	254 *S_s_z	63 *I_dh	44 *D_d
353 *S_ey_ea	349 *I_oh	198 *D_hh	283 *S_ax_uw	351 *I_oh	195 *D_s	250 *S_eh_ax	58 *I_s	43 *D_ih
323 *S_ih_iy	340 *I_hh	192 *D_l	264 *S_ax_ih	347 *I_ah	192 *D_hh	238 *S_uw_ax	52 *I_z	38 *D_y
319 *S_s_z	323 *I_ah	189 *D_er	261 *S_sp_d	320 *I_ay	181 *D_er	201 *S_eh_ae	46 *I_r	36 *D_l

Table 23: Output of top frequent errors for ChildEn corpus generated from 5-gram model

ASR & MAN			ASR & IDE			IDE & MAN		
Substitutions	Insertions	Deletions	Substitutions	Insertions	Deletions	Substituions	Insertions	Deletions
453 *S_t_sil	1447 *I_sil	1462 *D_r	477 *S_t_sil	1673 *I_sil	1181 *D_@bg	520 *S_i_ih	1361 *I_r	82 *D_hh
427 *S_d_sil	76 *I_ea	1294 *D_@bg	292 *S_ax_sil	95 *I_ea	675 *D_ax	319 *S_d_dh	355 *I_ax	42 *D_d
284 *S_g_sil	57 *I_l	672 *D_ax	266 *S_ih_iy	73 *I_hh	504 *D_r	291 *S_o_ax	205 *I_g	39 *D_t
277 *S_r_sil	55 *I_t	421 *D_l	265 *S_dh_sil	58 *I_t	399 *D_l	260 *S_s_z	188 *I_R	33 *D_k
253 *S_i_ih	53 *I_dh	341 *D_n	251 *S_g_sil	53 *I_l	362 *D_ih	250 *S_R_r	115 *I_hh	31 *D_n
249 *S_ax_sil	51 *I_hh	296 *D_ih	248 *S_ax_ah	51 *I_ao	348 *D_n	217 *S_e_ax	96 *I_o	19 *D_th
188 *S_ih_iy	47 *I_d	296 *D_ay	227 *S_d_sil	50 *I_ax	306 *D_ay	199 *S_a_ax	66 *I_i	16 *D_y
157 *S_b_sil	37 *I_z	259 *D_g	180 *S_ax_ea	43 *I_z	227 *D_t	151 *S_oh_ax	35 *I_oh	16 *D_ax
155 *S_R_r	35 *I_n	254 *D_o	177 *S_b_sil	43 *I_d	214 *D_z	126 *S_o_ow	35 *I_d	14 *D_l
151 *S_p_sil	31 *I_ax	235 *D_s	161 *S_n_sil	37 *I_f	209 *D_s	112 *S_u_uh	34 *I_v	13 *D_r
151 *S_i_iy	22 *I_r	226 *D_t	161 *S_ih_sil	35 *I_n	192 *D_iy	96 *S_e_ih	31 *I_a	10 *D_z

Appendix E Adapted lexicon

Table 24: Sample of the adapted lexicon

Word	Adapted transcription
SUPPORTING	s a p a o t i n g s a p a o t i n g g s a p a o t i h n g s a p a o t i h n g g s a p a o t i y n g s a p a o t i y n g g s a x p a o t i n g s a x p a o t i n g g s a x p a o t i h n g s a x p a o t i h n g g s a x p a o t i y n g s a x p a o t i y n g g s o p a o t i n g s o p a o t i n g g s o p a o t i h n g s o p a o t i h n g g s o p a o t i y n g s o p a o t i y n g g s o h p a o t i n g s o h p a o t i n g g s o h p a o t i h n g s o h p a o t i h n g g s o h p a o t i y n g s o h p a o t i y n g g
JAPANESE	j h a e p a n i y s j h a e p a n i y z j h a e p a x n i y s j h a e p a x n i y z j h a e p o n i y s j h a e p o n i y z j h a e p o h n i y s j h a e p o h n i y z
HAPPIER	a e p i a a e p i a R a e p i a r h h a e p i a h h a e p i a R h h a e p i a r

Appendix F Phone list of Ideal, Manual, & ASR

Remarks: **sp** -> Add sp (short pause) to word and syllable boundaries. **sil** -> Add sil (Silence) at the beginning and end of an utterance.

Table 25: ISLE phone list: Ideal, Manual & ASR

All phones	IDE		MAN		ASR	
a	-	-	-	-	8969	a
aa	782	aa	1022	aa	983	aa
ae	1923	ae	1791	ae	2137	ae
ah	2253	ah	1940	ah	1376	ah
ao	1019	ao	1105	ao	862	ao
aw	321	aw	304	aw	377	aw
ax	4545	ax	6215	ax	778	ax
ay	2639	ay	2430	ay	2979	ay
b	1814	b	1882	b	2062	b
ch	411	ch	381	ch	448	ch
d	3509	d	4102	d	5657	d
dh	2151	dh	1376	dh	-	-
e	-	-	-	-	1	e
E	-	-	-	-	18	E
ea	-	-	-	-	415	ea
eh	2654	eh	3144	eh	3053	eh
er	1780	er	1600	er	905	er
ey	2395	ey	2313	ey	1296	ey
f	1284	f	1407	f	1338	f
g	850	g	1116	g	1041	g
hh	902	hh	864	hh	602	hh
i	-	-	-	-	4801	i
ia	-	-	-	-	147	ia
ih	4587	ih	3315	ih	-	-
iy	2767	iy	3875	iy	2438	iy

j	-	-	-	-	9	j
jh	644	jh	564	jh	668	jh
k	3205	k	3127	k	3031	k
l	3037	l	3072	l	3279	l
m	1801	m	1803	m	1928	m
n	5610	n	5573	n	6189	n
ng	842	ng	827	ng	948	ng
o	-	-	-	-	98	o
O	-	-	-	-	27	O
oh	2571	oh	3054	oh	2173	oh
ow	1092	ow	985	ow	1129	ow
oy	159	oy	139	oy	173	oy
p	2433	p	2491	p	2634	p
r	4159	r	4517	r	-	-
R	-	-	-	-	3482	R
s	4976	s	5030	s	7226	s
sh	671	sh	684	sh	750	sh
sil	11683	sil	11676	sil	20160	sil
sp	17560	sp	17568	sp	-	-
t	7372	t	6996	t	6280	t
th	311	th	281	th	332	th
u	-	-	-	-	454	u
ua	-	-	-	-	36	ua
uh	476	uh	580	uh	-	-
uw	1121	uw	1518	uw	842	uw
v	1329	v	1245	v	1422	v
w	1684	w	1696	w	1837	w
y	711	y	693	y	709	y
z	1752	z	1596	z	-	-
zh	105	zh	144	zh	125	zh
z	1130	z	938	z	-	-

zh	261	zh	229	zh	264	zh
-----------	-----	----	-----	----	-----	----

Table 26: ChildEn phone list: Ideal, Manual & ASR

All phones	IDE		MAN		ASR	
a	-	-	523	a	2586	a
aa	442	aa	376	aa	511	aa
ae	628	ae	635	ae	594	ae
ah	658	ah	561	ah	663	ah
ao	497	ao	431	ao	606	ao
aw	301	aw	243	aw	275	aw
ax	2585	ax	1812	ax	440	ax
ay	732	ay	719	ay	748	ay
b	1249	b	1247	b	1249	b
ch	427	ch	453	ch	427	ch
d	1136	d	1437	d	1546	d
dh	678	dh	287	dh	-	-
dz	-	-	16	dz	58	dz
e	-	-	480	e	9	e
E	-	-	1	E	59	E
ea	320	ea	302	ea	283	ea
eh	866	eh	975	eh	912	eh
er	310	er	363	er	395	er
ey	609	ey	522	ey	577	ey
f	821	f	916	f	823	f
g	861	g	1063	g	958	g
hh	422	hh	453	hh	276	hh
i	-	-	724	i	2486	i
ia	294	ia	214	ia	233	ia
ih	2619	ih	1956	ih	-	-
iy	551	iy	465	iy	585	iy
j	-	-	10	j	-	-
jh	356	jh	308	jh	352	jh
k	1448	k	1432	k	1312	k

l	1721	l	1724	l	1721	l
m	1045	m	1040	m	1042	m
n	2096	n	2088	n	2095	n
ng	540	ng	518	ng	539	ng
o	-	-	673	o	185	o
O	-	-	5	O	38	O
oh	680	oh	924	oh	605	oh
ow	495	ow	289	ow	449	ow
oy	275	oy	259	oy	236	oy
p	951	p	951	p	950	p
r	1538	r	2701	r	-	-
R	-	-	452	R	1668	R
s	1290	s	1506	s	2356	s
sh	316	sh	324	sh	316	sh
sil	12834	sil	12833	sil	13734	sil
t	2341	t	2338	t	1874	t
th	292	th	175	th	286	th
ts	-	-	14	ts	-	-
u	-	-	323	u	311	u
ua	220	ua	101	ua	152	ua
uh	343	uh	280	uh	-	-
uw	528	uw	430	uw	532	uw
v	583	v	600	v	577	v
w	611	w	623	w	611	w
y	516	y	504	y	519	y
z	1130	z	938	z	-	-
zh	261	zh	229	zh	264	zh

References

- Ai, R. (2018). *Speech verification for computer assisted pronunciation*. PhD thesis, Saarländische Universitäts- und Landesbibliothek.
- Amdal, I., Johnsen, M. H., and Versvik, E. (2009). Automatic evaluation of quantity contrast in non-native Norwegian speech. In *International Workshop on Speech and Language Technology in Education*.
- Batliner, A., Blomberg, M., D’Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., and Wong, M. (2005). The PF_STAR children’s speech corpus. In *Ninth European Conference on Speech Communication and Technology*.
- Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., and Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. In *First International Conference on Spoken Language Processing*. First International Conference on Spoken Language Processing.
- Boersma, P. and Weenink, D. (2016). Praat: Doing phonetics by computer (Version 6.0.14). Retrieved from (last access: 29.04. 2018).
- Browning, S. (2004). Analysis of Italian children’s English pronunciation. *Report contributed to the EU FP5 PF STAR Project*.
- Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., and Gyawali, B. (2018). Automated Scoring of Nonnative Speech Using the SpeechRaterSM v. 5.0 Engine. *ETS Research Report Series*, 2018(1):1–31.
- Chen, N. F. and Li, H. (2016). Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–7.
- Derwing, T. M. and Munro, M. J. (2005). Second Language Accent and Pronunciation Teaching: A Research-Based Approach. *TESOL Quarterly*, 39(3):379.

- Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., and Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly*, 5(2):160–167.
- Errattahi, R., Hannani, A. E., and Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37.
- Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning and Technology*, 2(2):62–76.
- Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.
- Fant, G. (1973). *Speech sounds and features*. The MIT Press.
- Franco, H., Neumeyer, L., Kim, Y., and Ronen, O. (1997). Automatic pronunciation scoring for language instruction organization. In *1997 IEEE international conference on acoustics, speech, and signal processing*, pages 1471–1474. IEEE.
- Franco, H., Neumeyer, L., Ramos, M., and Bratt, H. (1999). Automatic detection of phone-level mispronunciation for language learning. In *Sixth European Conference on Speech Communication and Technology*.
- Fu, J., Chiba, Y., Nose, T., and Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, 116:86–97.
- Gretter, R., Matassoni, M., Allgaier, K., Tchistiakova, S., and Falavigna, D. (2019). Automatic Assessment of Spoken Language Proficiency of Non-native Children. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2019-May, pages 7435–7439. IEEE.
- Hönig, F., Batliner, A., Weilhammer, K., and Nöth, E. (2010). Automatic assessment of non-native prosody for english as L2. In *Speech Prosody 2010–Fifth International Conference*.
- Huensch, A. (2019). The pronunciation teaching practices of university-level graduate teaching assistants of French and Spanish introductory language courses. *Foreign Language Annals*, 52(1):13–31.

- International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.
- Jurafsky, D. and Martin, J. H. (2018). N-gram language models. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.
- Kato, H., Harada, S., Kitade, T., and Shiga, Y. (2020). Multilingualization of Speech Processing. In *Speech-to-Speech Translation*, pages 1–20. Springer.
- Kawai, G. and Hirose, K. (1998). A CALL system using speech recognition to teach the pronunciation of Japanese tokushuhaku. In *STiLL-Speech Technology in Language Learning*.
- Kim, Y., Franco, H., and Neumeyer, L. (1997). Automatic pronunciation scoring of specific phone segments for language instruction. In *Fifth European Conference on Speech Communication and Technology*.
- Kruk, M. (2012). Using online resources in the development of learner autonomy and English pronunciation: The case of individual learners. *Journal of Second Language Teaching and Research*, 1(2):113–142.
- Madikeri, S., Dey, S., Motlicek, P., and Ferras, M. (2016). Implementation of the standard i-vector system for the kaldi speech recognition toolkit (No. REP_WORK). Technical report, Idiap.
- Maqsood, M., Adnan Habib, H., Nawaz, T., and Zeeshan Haider, K. (2016). A Complete Mispronunciation Detection System for Arabic Phonemes using SVM. *IJCSNS International Journal of Computer Science and Network Security*, 16(3):30.
- Markov, A. (1913). Essay of a statistical research on the text of the novel “Eugene Onegin” illustrating the linkage of chain tests (example of a statistical investigation of the text of Eugene Onegin illustrating the dependence between samples in chain). *Izvestia Imperatorskoi Akademii Nauk (Bulletin of the French Academy of Sciences of St. Petersburg)*, 7(153-162):209.
- McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *System*, 57(February):25–42.

- Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., and Souter, C. (2000). The ISLE corpus of non-native spoken English. In *2nd International Conference on Language Resources and Evaluation, LREC 2000*, volume 2, pages 957–964. European Language Resources Association.
- Miao, Y., Zhang, H., and Metze, F. (2015). Speaker adaptive training of deep neural network acoustic models using i-vectors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1938–1949.
- Minematsu, N. (2004). Pronunciation assessment based upon the phonological distortions observed in language Learners’ utterances. In *8th International Conference on Spoken Language Processing, ICSLP 2004*, pages 1669–1672.
- Neri, A., Cucchiaroni, C., Strik, H., and Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer assisted language learning*, 15(5):441–467.
- Neri, A., Cucchiaroni, C., and Strik, W. (2003). Automatic speech recognition for second language learning: how and why it actually works. In *In Proc. ICPhS*, pages 1157–1160.
- Nicolao, M., Beeston, A. V., and Hain, T. (2015). Automatic assessment of English learner pronunciation using discriminative classifiers. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5351–5355. IEEE.
- Peabody, M. A. (2011). *Methods for pronunciation assessment in computer aided language learning*. PhD thesis, Massachusetts Institute of Technology.
- Power, K., Morton, R., Matheson, C., and Ollason, D. (1996). The Graphvite book 1.1. *Entropic, Cambridge*.
- Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286.
- Raux, A. and Kawahara, T. (2002). Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning. In *7th International Conference on Spoken Language Processing, ICSLP 2002*, pages 737–740.

- Ryu, H., Hong, H., Kim, S., and Chung, M. (2017). Automatic pronunciation assessment of Korean spoken by L2 learners using best feature set selection. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*.
- Strik, H., Truong, K., de Wet, F., and Cucchiaroni, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51(10):845–852.
- Tao, J., Ghaffarzadegan, S., Chen, L., and Zechner, K. (2016). Exploring deep learning architectures for automatically grading non-native spontaneous speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6140–6144. IEEE.
- Truong, K., Neri, A., De Wet, F., Cucchiaroni, C., and Strik, H. (2005). Automatic detection of frequent pronunciation errors made by L2-learners. In *9th European Conference on Speech Communication and Technology*, pages 1345–1348.
- Van Doremalen, J., Cucchiaroni, C., and Strik, H. (2009). Automatic detection of vowel pronunciation errors using multiple information sources. In *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, pages 580–585. IEEE.
- Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., Binder, J. R., Men, W., Gao, J. H., and Bi, Y. (2018). Organizational principles of abstract words in the human brain. *Cerebral Cortex*, 28(12):4305–4318.
- Witt, S. and Young, S. (1997). Computer-assisted pronunciation teaching based on automatic speech recognition. *Language Teaching and Language Technology Groningen, The Netherlands*.
- Witt, S. and Young, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3):95–108.
- Witt, S. M. (1999). *Use of speech recognition in computer-assisted language learning*. PhD thesis, University of Cambridge Cambridge, United Kingdom.
- Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 6:1–8.

Xiao, W. (2018). Automatic Speech Recognition and Pronunciation Training. In *2018 2nd International Conference on Education, Economics and Management Research (ICEEMR 2018)*. Atlantis Press.

Zechner, K., Higgins, D., Xi, X., and Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.