UPPSALA
UNIVERSITET

Low-Resource CAPT: An Irish Perspective

Peter Cady

Uppsala University
Department of Linguistics and Philology
Master Programme in Language Technology
Master's Thesis in Language Technology, 30 ECTS credits

September 17, 2025

Supervisors:
Johan Sjons, Uppsala University
Jim O' Regan, KTH Royal Institution of Technology

Abstract

[1] summarize the motivation
[2] problem overview
[3] experiment
[4] main findings and contributions.

Contents

Preface 4

Acknowledgements 5

1 Introduction 6

2 Background 8

2.1 Irish & The Predicament of Minority Languages in Natural Language Processing (NLP) . . . 8

2.2 From Text to Speech: Approaches to Low-Resource Scenarios . . . 12

2.3 Computer-Assisted Language Learning (CALL) & Computer-Assisted Pronunciation Training (CAPT) . . . 13

2.4 Pronunciation & Mispronunciation Detection and Diagnosis (MDD) . 14

2.5 Speech and automatic speech recognition (ASR) . . . 15

3 Related Work 17

3.1 end-to-end (E2E) automatic speech recognition (ASR) for Mispronunciation Detection and Diagnosis (MDD) . . . 17

3.2 ASR Ensembling Strategies . . . 17

3.3 Data Augmentation with Synthetic speech . . . 19

4 Methods & Materials 20

4.1 General Experiment Setup . . . 20

4.2 Experiment 1: Monolingual Model Ensembling . . . 21

4.2.1 Data . . . 21

4.2.2 Experimental Setup . . . 22

4.3 Experiment 2: Synthetic Mispronunciation Data . . . 22

4.3.1 Data . . . 22

4.3.2 Experimental Setup . . . 23

4.4 Evaluation . . . 23

4.4.1 Metrics . . . 23

4.4.2 Data . . . 23

4.4.3 Skyline Comparison . . . 24

5 Results 25

6 Discussion 26

7 Conclusions 27

8 Ethical Considerations 28

9 AI Tools 29

Preface

Acknowledgements

Thank my partner Thank Johan I would especially like to thank Jim O' Regan for all his thought-provoking insights, encouragement, and advice, which were crucial for me getting my bearings in a project which pushed me to grow well beyond my limits. The process was deeply meaningful to me. Thank you. Thank Greg I would like to thank my Irish professor Gregory Darwin for his commitment and engagement in promoting Irish literature and the Irish language Thank examiner

1 Introduction

[5] Lead work here. Outline modern context of minority languages in nlp. Introduce Irish context in this framework. revisit introduction, try to lead with an impactful framing of the motivation of the work. Language competency is crucial for securing opportunities in the workplace as well as for accessing services and exercising rights in society at large. As our surroundings become ever more digitally interconnected, these interactions are increasingly mediated by language technologies, for both good and ill. Such technologies are often identified as promising tools to promote language use and language learning, but are also implicated in the coalescence of online interaction around a few dominant languages such as English, complicating the glowing promises often offered by their proponents. Low-resource language communities struggle to keep pace with the most recent technological advancements given the relative scarcity of resources they are faced with, necessitating approaches tailored to these limitations if the true promise of cutting-edge language technologies is to be realized for these groups most in need. The momentum of the pre-training/fine-tuning paradigm within ASR and Machine Learning (ML) more generally in recent years offers a glimmer of hope to those working toward making accessible tools for such communities, enabling the use of more plentiful data to support tools for languages facing varying levels of resource scarcity.

[6] Reiterate need for a solution, lead into research questions. Promoting speech technologies for language learning could be of particular benefit for languages such as Irish which struggle to propagate native models of speech effectively to motivated learners outside of traditionally Irish-speaking areas. Through Computer-Assisted Pronunciation Training (CAPT) applications built with careful use of the aforementioned, we might find the scaleable, resource con

[7] RQ's To this end, in this work we explore two potential avenues to overcoming the data limitations faced by Irish and other languages like it: one, by using the data we *do* have by harnessing readily available monolingual data in a model ensemble; and the other, by imitating the data we *wish* we had with text-to-speech (TTS)-generated learner approximations to train a model with. These approaches are formulated concretely as:

1. To what extent can a resource-conscious ensemble of monolingual ASR models (Irish and English) approach the performance of a high-resource upper-bound model trained on fully annotated mispronunciation data (L2-ARCTIC)?

2. To what extend can a model trained on TTS-generated synthetic mispronunciation data approach the performance of a high-resource upper-bound trained on fully annotated learner mispronunciation data (L2-ARCTIC)?

[8] Outline key goals of thesis, give a "road map" for what's to come. In this thesis we explore the feasibility of two primary methods of overcoming the scarcity of phonetically annotated second language (L2) Irish learner data for Mispronunciation Detection (MD) applications. One approaches the problem with ensembles of monolingual ASR models for which data scarcity less acute, and another using a schema of data generation by leveraging established TTS systems to approximate learner speech: a potential low-cost alternative to large-scale phonetic annotation. By exploring these approaches, we aim to illuminate possible ways forward for low-resource language communities interested in developing automated technologies for language learning and pronunciation training.

2 Background

[9] Short overview of background section

In the following section we will outline the context motivating our current work with relation to Natural Language Processing (NLP) and ASR research for minority languages. We begin by summarizing the state of minority languages in current research, highlighting some promising trends as well as some thusfar recalcitrant problems hindering more equal access to the benefits of our time's rapid advances in language technology. We then proceed to outline the more general research space of the current work: Computer-Assisted Language Learning (CALL) and CAPT before digging into the core task of CAPT systems: MD and MDD. We conclude with a technical overview of modern ASR systems used for such tasks, leaving the specific efforts of researchers to overcome data limitation inherent to low-resource languages in the next section, get chapter and section references working

2.1 Irish & The Predicament of Minority Languages in Natural Language Processing (NLP)

[10] introduce need for general-use systems.

Developing language technologies that can scale beyond the language they are designed for is no new goal for NLP research. The value of such a property is apparent: transferring an existing system seamlessly to another language could potentially save significant resources for language communities without the ability to fund such system development themselves. Actually achieving this goal in practice, however, is no simple endeavor, typically requiring some level of linguistic awareness which is all-to-often lamentably absent (see Bender, 2011; Hedderich et al., 2021; Joshi et al., 2021, inter alia). For example, success in applying word-based n-gram approaches to a language rests on the language's level of inflectional morphology: the lower the better. Higher morphological complexity together with variations in word order raise data sparsity problems which n-gram approaches rooted in English struggle to handle (Bender, 2011). This should come as no surprise, given the relatively fixed word order and low levels of inflectional morphology present in English, but it illustrates a need for caution: systems developed for a given language may make implicit assumptions about language structure which do not generalize well.add concrete example to illustrate? maybe exemplify more directly from above reference Linguistic typology can provide important clues as to what features are shared between the original development language and possible languages of extension for a system. Information of this kind has, for many of the worlds languages, already been gathered by linguists. Perhaps the most renowned database of typological information is the World Atlas of Linguistic Structure (WALS) (Dryer et al., 2024), a free, online resource currently boasting 152 chapters with detailed descriptions of 192 linguistic features spanning over 2,600 of the world's languages. By explicitly mobilizing linguistic knowledge already painstakingly gathered by linguistic typologists, we can identify where languages agree, where they differ, and hopefully identify some of the implicit, ungeneralizable assumptions that have hindered past efforts to create more accessible language technologies.

[11] language disparity in lang tech. Despite claims of language-agnostic systems often touted by proponents of emerging Artificial Intelligence (AI) systems, these have often also fallen well short of such promise (Bender, 2011)add some examples of where to read further + inter alia. The overwhelming majority of the world's languages still have no footprint in emerging language technologies (Joshi et al., 2021). In the past, building neural network (NN)-based language technologies demanded immense quantities of labeled data: a high bar of entry to the language communities of those languages with limited if any access to such resources, and an ongoing issue which continues to stimulate a body of research dedicated to overcoming it (see Maguéresse et al., 2020, for an overview). For languages where data availability is no obstacle, research and development can proceed unfettered by the prohibitive cost of curating datasets from scratch. As our daily lives grow increasingly integrated with the digital realm, language communities without the same support are obliged to switch to more digitally dominant languages (often English) to gain access to these new resources, narrowing the opportunities to engage with resources and services through the medium of their community language (Ní Chasaide et al., 2019). A particularly sobering taxonomy illustrating the states of languages facing such disparities is formulated by Joshi et al. (2021) and reproduced in Table 2.1, which outlines the challenges faced by languages in resource terms in the digital space, and how dominant a small group of languages are within it. Those at the bottom of the data availability scale at level 0, 'The Left-Behinds' have virtually no data of any kind available to support language technology development. those in the middle at level 2 or 3 have some resources, net presence, and/or research communities supporting them, but critically lack in substantial quantities of the labeled data typically required for cutting-edge NLP tools. At the top, languages like English reap the lion's share of overall investment and development. Of particular note for the privileged few that find themselves at the top of the heap is their typological similarity, being drawn as they are from a few dominant language families (and even dominant branches within these larger families). This state of affairs constitutes a sort of typological echo-chamber for the cutting edge of NLP developments, a point which we will return to shortly.

[12] set the Irish case in this context, outline challenges currently undertaken by developers

Despite some advantages not afforded other minority languages, Irish still finds itself struggling to maintain a footing in the the digital realm, placed by Joshi et al. (2021) in class 2 of the taxonomy in Table 2.1. It enjoys ongoing investment by the Irish state, nominal status of Irish as the first national language of the Republic of Ireland, and research dedicated to its promotion (e.g. through the ABAIR initiative dedicated to developing speech technologies for Irish, see Chasaide et al., 2017). At the same time, it is a typological outlier in several respects: it is a verb-initial language with relatively complex inflectional morphology, and a distinct (though still Latin-based) orthography. Features like these put Irish at odds with many languages in the high-resource echo-chamber, complicating the ability to leverage cutting-edge models to linguistic features with no representation in a model's training data. Furthermore, though we have treated Irish as a single entity thus far, an important complication reveals itself in the discontiguous nature of the Irish-speaking areas, referred to as *the Gaeltacht* or *the Gaeltachtaí*(pl.). Each of the three main areas (i.e. Ulster, Connacht, and Munster 12) speak markedly distinct varieties of Irish, necessitating labeled data from each variant if these groups are to be adequately serviced by new technologies (Ní Chasaide et al., 2019). In the face of such limitations, today's data-hungry tools simply cannot be expected to perform to the same level on languages like Irish without the necessary resources. It should be noted that the pretraining/finetuning paradigm of recent massive multilingual models does mitigate this demand for data somewhat

Class Descriptions	Example Languages	% of total languages
0 The Left-Behinds: Virtually ignored in language technology. Exceptionally limited resources available, even with respect to unlabeled data.	Dahalo, Bora	88.17%
1 The Scraping-Bys: Some unlabeled data. With organized promotion and data collection, there is hope for improvement in coming years.	Fijian, Navajo	8.93%
2 The Hopefuls: Limited labeled data. Support communities help these languages survive, and there is promise for NLP tools in the near term.	Zulu, **Irish**	0.76%
3 The Rising Stars: Strong web presence and thriving community online. Lacking labeled data. Good potential for NLP tool development for these languages.	Indonesian, Hebrew	1.13%
4 The Underdogs: Much unlabeled data, and less but still significant labeled data. Dedicated investment from NLP communities.	Russian, Dutch	0.72%
5 The Winners: Dominant online presence with massive investment and resources.	English, German	0.28%

Table 2.1: Data availability & status taxonomy of languages adapted from work by Joshi et al. (2021).

by leveraging unlabeled cross-lingual data, reducing the need for labeled data in the language finetuned to (Hedderich et al., 2021; Joshi et al., 2021; Ranathunga et al., 2021), but for other languages without even minimal labeled data to their name, this is a small comfort.

[13] problems to be solved for Irish speakers Though the hurdles facing the Irish language in the digital sphere are a relatively recent concern, the issues facing it in the real world are anything but. It is currently classified by UNESCO as being *definitely endangered* (*Atlas of the World's Languages in Danger* 2010), following centuries of varying rates of contraction due to encroachment by English give some more explanation of the history, handle with care. Irish survives as a community languages in the aforementioned Gaeltacht areas, though even there it is estimated that only 24% of inhabitants speak Irish on a daily basis (Ní Chasaide et al., 2015). Despite the state of Irish as a first language (L1), it is comparatively strong as a L2 (Broin, 2014). A growing number of parents seek Irish-medium education for their children outside the Gaeltacht, and immersive summer courses remain popolar among adults looking to learn or reconnect with the language. This encouraging L2 engagement intersects with thorny issues of supply, however, as many of the teachers are not themselves native speakers with an accompanying native grasp of the structure and sound of the language (Ní Chasaide et al., 2015, 2019). This limited native speaker model for L2 speakers is particularly problematic for teaching pronunciation, complicating the acquisition of some sound contrasts critical to disambiguating the Irish grammar. Perhaps chiefly among these, contrasts between the secondary articulation of consonants into *palatalised* and *velarised* variants play an instrumental role in a number of grammatical functions, such as in the formation of certain plurals and genitive marking (Broin, 2014; Gabriele, n.d.; Snesareva, 2016; Stenson, 2020). Since the Roman alphabet doesn't provide symbols for this distinction, Irish orthography marks it via adjacent letters as seen in give sub-table references of Table 2.2: so-called *slender* vowels ('i' and 'e') around a consonant denote *palatalisation*, while *broad* vowels ('a', 'o', and 'u')

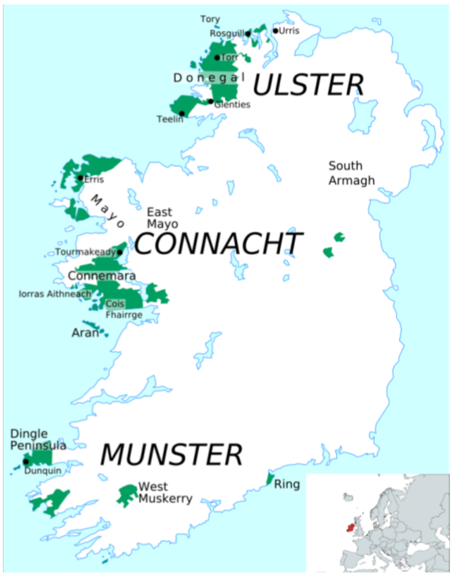

Figure 2.1: Area of the Gaeltachtaí (Irish-speaking areas of Ireland) colored in green. The original uploader of the map of Ireland was Angr at English Wikipedia. - Transferred from en.wikipedia to Commons., CC BY-SA 3.0. European map was created with mapchart.net

mark *velarisation*[1] (Stenson, 2020). Mutation effects are another pervasive element of the Irish grammar relying on sound alterations, the most common of which are *lenition* and *eclipsis*. Lenition, traditionally termed *séimhiú* (/ˈʃeːvʲuː/), is commonly marked with an 'h' following the lenited consonant as seen on Table 2.2, and originally denoted a weakening in the manner of articulation, though the relationship between consonants and their lenited versions is less immediately apparent now for some consonants (Stenson, 2020). Eclipsis, traditionally *úru* (/ˈʊɾˠuː/), involves replacing the original consonant with a nasalized or voiced version, and is denoted by appending the new sound character before the consonant being eclipsed2.2 add sub-table refs and double check if this description covers all the bases. These and other structural underpinnings of the language can have far-reaching implications for intelligability if not adequately mastered by students. Indeed, a study undertaken by Broin (2014) reveals that realisations of phonological details like those above have error rates of over 50% on average for urban speakers, with some as high as 82%—a stark departure from their gaeltacht counterparts which lie below 10%. Providing wider access to better native models of pronunciation—not to mention native models of morphology—could do much to close this gap, making mutual intelligability between gaeltacht and urban speakers more attainable.

Table 2.2: Examples use of a. consonant velarisation & b. palatalisation c. séimhiú & d. úru. Adapted from Ní Chasaide et al. (2015)

	Orthographic	the International Phonetic Alphabet (IPA)	Translation
a.	bád	/bˠaːdˠ/	'boat'
b.	báid	/bˠaːdʲ/	'boats'
c.	do bhád	/dˠɔ waːdˠ/	'your boat'
d.	ár mbád	/ɛɾʲ mˠaːdˠ/	'our boat'

[1] For initial consonants, the following vowel plays the determining role, for final consonants, the preceding vowel, and for medial consonants, the vowels flanking it.

Gabriele (n.d.) english influence on palatalization and velarization Snesareva (2016) how does english influence Irish spoken by dublin bilinguals? Broin (2014) differences of irish between cities and gaeltacht Ní Chasaide et al. (2015) documentation of Irish with speech tech

Magueresse et al. (2020) survey of low resource methods in NLP Wu et al. (2021) motivation for phone-based recognition for MDD instead of scoring pronunciations (like GOP)

2.2 From Text to Speech: Approaches to Low-Resource Scenarios

With the rise of NN-based language technologies, the data-hungry nature of such tools underscores the urgency of addressing the kind of resource disparities outlined above. Making such tools accessible to languages without the same strong data foundations as English is an active area of ongoing research, though even within popular languages such as English, non-standard domains and tasks types can constitute low-resource areas which lack suitable quantities of training data. These data disparities can be categorized along several dimensions, such as those proposed for NLP by Hedderich et al. (2021) as: availability of *task-specific labeled data* for the target language or domain, availability of *unlabeled* language- or domain-specific data, or the availability of *auxiliary* data. This latter kind of data is diverse, as it is data not directly labeled for the task at hand, but which can still be indirectly useful, from labels specific to another language/domain, to knowledge bases such as entity lists, or automated labels from Machine Translation (MT) systems (Hedderich et al., 2021).

[14] overview of low resource approaches To address the different dimensions of resource scarcity outlined above, various approaches have been developed which Hedderich et al. (2021) splits broadly into those which *generate additional labeled data*, and those employing Transfer Learning (TL). Faced with limited gold-standard annotated data, researchers employ strategies of the former type to (semi-)automatically produce labeled alternatives. These strategies can be themselves broadly grouped as *data augmentation*, where task-specific labeled data is used to make more labeled data, such as with Back-Translation for MT where a target-to-source translation model is used to obtain a synthetic parallel corpus from a monolingual target corpus (Ranathunga et al., 2021), and *distant supervision* which produces labels for existing unlabeled data, for example in Cross-Lingual Annotation Projection where a task-specific classifier is trained for a high-resource language, then projected onto text from a low-resource language using a paralell corpus. For TL, in contrast, instead of creating or extending task-specific training data, the focus lies on reducing the need for such data by leveraging models or learned representations from other languages/domains. This approach has been particularly successful in in recent years with the advent of models like BERT (Devlin et al., 2019) and Wav2Vec2 (Baevski, Zhou, et al., 2020) which are *pre-trained* on vast quantities of unlabeled data to then be *fine-tuned* for specific downstream tasks. This pretraining/fine-tuning paradigm can be particularly advantageous for languages or domains where labeled data is limited.

[15] Connect these general trends to applicability in Automatic Speech Recognition Although these strategies are commonly employed in NLP, analogous trends can be found in computer vision as well as speech to tackle similar limitations in data. TL and the aforementioned pretraining/fine-tuning paradigm has made strides with self-supervised models like Wav2vec2, outperforming previous state-of-the-art ASR models with 100 times less data, starkly reducing the demand for labeled speech (Baevski, Zhou, et al., 2020). Alongside TL, ensemble methods have also emerged as a promising tool for low-resource contexts. Here, specialist models with complementary attributes can be combined to perform better than any one of its constituents for novel tasks (Arunkumar et al., 2022; L. Deng and Platt, 2014; Fiscus, 1997; Gitman et al., 2023, inter alia). Various methods of data augmentation and data synthesis are also prevalent to improve ASR performance for low-resource languages, including voice transformations, where noise or other alterations are introduced to recordings to extend existing data, or generating synthetic audio data samples with a TTS system to bolster training data when authentic speech corpora are lacking (Bartelds et al., 2023; D. Zhang et al., 2022).

Hedderich et al. (2021) survey of low resource NLP methods Magueresse et al. (2020) survey of low resource methods in NLP

2.3 Computer-Assisted Language Learning (CALL) & Computer-Assisted Pronunciation Training (CAPT)

Despite efforts to the contrary, the rise of digital technologies is often implicated in the acceleration of already precpitous rates of decline for endangered languages dig up reference to strenghten this point. However, it is a trend that cuts both ways, as the same technologies that squeeze certain languages out of the digital realm are also making space for communities of language learners to come together towards their common goal through CALL platforms and massive open online course (MOOC)s. The increased presense of technology both in and outside the classroom brings with it broad implications for traditional pedgagogy, enabling more autonomous and flexible modes of learning for students (Spolsky and Hult, 2008)alter bib entry to reflect book chapter, not whole book particularly for learners looking to autonomously improve their pronunciation via CAPT systems. This technological shift has increased the financial viability of courses for endagered or otherwise less commonly taught languages, allowing teachers to draw from a more geographically dispersed enrollment pool and provide courses otherwise impossible to offer. Despite this potential, tension between the technology and pedagogy underpinning CALL and CAPT systems remains a well-documented issue, and their increasing prevalence in the modern language learning landscape has renewed calls for greater collaboration between pedagogical and technical experts when designing these systems (Rogerson-Revell, 2021). This tension also crosscuts the aforementioned struggles of low-resource languages in making the most of recent technological developments. Although the scope of this work doesn't evaluate the effectivness of any particular pedagogical application, we recognize the potential of CAPT systems and the need for pedagogical and linguistic awareness in their design to more fully deliver on their promise. To that end, we endeavor in this work to contribute with linguistically aware ASR strategies which can hopefully help bridge the gap between cutting-edge language technology and the needs minority language communities, supporting pedagogically sound feedback to learners.

The proliferation of CALL software for self-study such as Duolingo, Babbel, and Rosetta Stone proliferate, has renewed interest in the role of immediate and personalised feeback for pronunciation training in CAPT systems. Recent research indicates that language learners may need explicit and targeted feedback on their pronunciation in order to improve (Bajorek, 2017), despite the long-held understanding that error correction typically does not meaningfully influence acquisition (Krashen, 1984). Perhaps as a consequence of this established view, explicit pronunciation training has been notably absent from language classrooms in recent decades, and many of these

CALL platforms carry on this legacy with binary right-or-wrong feedback mechanisms that do not make use of the potential for more effective, targeted feedback (Bajorek, 2017). The feedback is also frequently unexplained, making such binary judgements about pronunciation quality opaque to the student and thus more difficult to act upon. The ideal individualized, explaination enriched feedback would normally carry a steep price tag for the student of an individual tutor, say, but CAPT systems can potentially lower this barrier of entry considerably by automating the same kind of undivided feedback in a one-to-many form scalable to many geographically disperesed students at once.

2.4 Pronunciation & Mispronunciation Detection and Diagnosis (MDD)

To realize the often untapped potential of CAPT and lay the groundwork for actionable feedback, the CAPT system must be able to identify deviations in student pronunciations from the target pronunciation and determine how it differs in ways interpretable to the student. To do this, we must start by specifying what we mean by pronunciation. In broad strokes, the human speech apparatus may be seen as a collection of subsystems which can emit signals over parallel channels (Engstrand, 2004, p. 173). Although speech is a continuous signal, we can map symbols of discrete speech sounds–phones–onto subsegments of this signal. These discrete units—the vowels and consonants that constitute words—are *segmental* features of speech. The prosody, intonation, stress, and other such elements of speech can be seen as superimposed on these segmental features, and are thus termed *suprasegmental* features of speech (Engstrand, 2004). Human cultures have an array of strategies for representing speech in written form, ranging from logographic writing systems like Chinese where one symbol represents one word, to different varieties of sound-based systems such as syllabic for Japanese hiragana or katakana, alphabetic like the Roman alphabet used here, or consonantal as in Semitic scripts (Jurafsky and Martin, 2025). Many, though by no means all, approaches to MD within CAPT focus on recognizing these segmental features in speaker recordings and comparing their phonetic transcriptions to a canonical pronunciation of speech for the target word (Korzekwa et al., 2022). For our purposes, we will narrow our investigation of MD to first focus on detection of these segmental features, representing pronunciation as strings of phones using the IPA standard of phonetic notation. Though suprasegmental features can also play a crucial role in disambiguating meaning, investigating them is beyond the scope of this thesis.

Approaches to segmental recognition for MD can be broadly divided into two approaches: those which align speech signals to canonical segment sequences, and those which first extract segments—generally phonemes or phones—from speech, then aligning the extracted sequence of segments to a canonical sequence for comparison (Korzekwa et al., 2022). The process of aligning speech to text which characterizes the former approach is termed *forced-alignment* and thus approaches to segment recognition with that focus are likewise termed forced-alignment approaches. Classic pronunciation scoring generally focused on these forced-alignment approaches, which typically relied on log-likelihood scores and log-posterior scores derived from Hidden Markov model (HMM)-based ASR model outputs (S. M. Witt, n.d.). The latter of these scores soon became the de facto standard due to its higher correlation with human assessments. Building further on this development, S. M. Witt (2000) introduced the widely used goodness of pronunciation (GOP) measure of pronunciation quality, which could be compared against a threshold value to determine how well the speaker's pronunciation matches the canonical model pronunciation (S. Witt and Young, 2014). Although GOP and other posterior-based scores continue to be used for MD with modern ASR architectures (Gong et al., 2022; Parikh et al., 2025, inter alia), with the recent advent of transformer-based E2E models, researchers have also pursued sequence alignment as an alternative to circumvent the need for phone-level forced-alignment of speech signals to reference transcriptions (Leung et al., 2019; Lo et al., 2020).

To detect mispronunciations at the segmental, we confine ourselves to the latter family of approaches outlined above: obtaining phones as an intermediate step before comparing to canonical strings. Detection of articulatorily meaningful units of speech could lay the ground for more informative feedback in any future CAPT extension, particularly in comparison to binary assessment strategies of simply identifying correct or incorrect pronunciation, but this focus introduces its own set of challenges. Recognizing phones requires fine-grained annotations at that level, which are difficult to obtain, particularly for L2 speech (D. Zhang et al., 2022). This framing of the MD problem is a data-scarce task to pursue, but the modern innovations in ASR architectures leveraging the aforementioned pre-training/fine-tuning paradigm could make it an increasingly feasible one, even for low-resource languages.

"With deep neural networks (DNNs), GOP is derived from posterior probabilities using the negative log of the mean softmax output over aligned frames" (Parikh et al., 2025)

Peng et al. (2021) Wav2vec2 for MDD (important) Peng et al. (2022) gating strategy (ignore irrelevant parts in transcription) and constrastive loss to reduce objective gap between phoneme recognition and MDD Shahin and Ahmed (2024) phonological-level MDD (articulatory focus) True acceptance rate, false rejection tc Stanley and Hacioglu (2012) difference in L1 dependent models vs baseline. how to introduce non native acoustic features. (variant of min phone error training that optimizes on maximizing discriminability between confusable phonetic units in nonnative acoustic space.) X. Xu et al. (2021) wav2vec2 for MDD

(Bartelds et al., 2023) survey of low-resource ASR (Besacier et al., 2014) survey of low-resource ASR

2.5 Speech and automatic speech recognition (ASR)

[16] general overview of ASR The critical technology underpinning segmental MD approaches is the ASR system transcribing spoken words into writing. ASR research spans many decades, having matured from its origins as a barely useful method of interface to the point where speech is the primary modality of interaction for popular forms of human-machine communication today (Yu and L. Deng, 2015). Some simpler tasks have been long since solved, such as simple yes-no recognition or digit recogniton (i.e. 0-9). Transcribing strings of phonemes or words like today's ASR models however, is far more complex, with many tens of thousands of possible vocabulary items to predict (Jurafsky and Martin, 2025). In recent decades, however, even these tasks have proven tractable under controlled conditions, as evidenced by the popularity of different commercial voice-controlled digital assistants such as Amazon's Alexa[2], Apple's Siri[3], or Google Assistant[4]. Modern ASR need to be robust to more than just vocabulary size, however. Jurafsky and Martin (2025) highlights other consequential dimensions of speech variability such as who the speaker is talking to; *read*

[2] https://alexa.amazon.com

[3] https://www.apple.com/siri/

[4] https://assistant.google.com/

speech such as that produced when humans talk to machines, is far easier to recognize than spontaneous *conversational speech* between two humans. The quality of the signal can also be a factor in the ease of recognition: speech recorded under optimal conditions in a recording studio will of course be easier to recognize than someone speaking at a distant microphone on a busy street. Finally—and perhaps most importantly for the current work—variation in speaker classes can impact robustness of an ASR when the speaker class lies outside what the system was trained on. Speaker class includes differences in *accent* or *dialect*, as is most relevant for our purposes, but also by age and sex/gender. Creating systems robust to all of this variability is by no means a solved problem, despite the notable progress in recent decades.

[17] old asr models, hmm gnn Prior to the turn of the century, HMM-Gaussian Mixture Model (GMM)-based models were the de facto standard for ASR, typically consisting of four main components: Signal processing and feature extraction, acoustic model, language model, and hypothesis search (Yu and L. Deng, 2015). These systems convert of an acoustic waveform into feature vectors by a signal processer, which are then combined by a decoder with a dictionary and language model or grammar network into a recognition network. Using this network, one can calculate the most likely word sequence given the waveform of the speech input (S. M. Witt, 2000) revisit this description. A number of popular toolkits have cropped up to support development of these systems, such as Kaldi (Povey et al., n.d.), and later ESPnet (Watanabe et al., 2018) as E2E architectures became more prevalent. These provided useful recipes to create reproducible systems for an array of ASR use cases. While these toolkits are still in use, the rise of multi-purpose ML libraries like Pytorch (Paszke et al., 2019) and Hugging Face (Wolf et al., 2020) have provided a simplified alternative to training modern ASR systems[5]

[18] Description of Wav2vec2 model Conneau et al. (2020) Wav2vec2 XLSR Baevski, Schneider, et al. (2020) (vector quantized) wav2vec2 with learned discrete representations Following the rise of Transformers in ML generally, transformer-based ASR models such as *wav2vec 2.0* have become increasingly popular for ASR tasks, including extracting phonemes from raw acoustic data. Similar to other E2E alternatives, it is conceptually simpler than other leading models, while its self-supervised framework achieves competitive performance while reducing the need for labeled data. After pre-training on unlabeled data, it can be fine-tuned on a much smaller set of labeled data, achieving state-of-the-art results in scenarios where labeled data is scarce Baevski, Zhou, et al. (2020). It consists of three main components: a Convolutional Neural Network (CNN)-based *feature encoder*, a *Transformer*-based network, and a *quantization module*. The feature encoder takes raw audio as input and outputs latent speech representations to be used both by the Transformer network and the quantization module. The Transformer captures contextual information about the encoder output, while the quantization module divides the encoder output into discrete speech representations.adapt diagram from wav2vec2 paper and explain the components properly. **graves2006connectionistempty citation** ctc loss

[19] motivation for wav2vec2 xlsr The ability to pre-train on freely available unlabeled data makes Wav2Vec 2.0 a natural choice for low-resource scenarios where one cannot count on abundant labeled data to train a model from the ground up. By leveraging the benefits afforded by the pre-training/fine-tuning paradigm, we can make the most of data limited both by domain and language. not done, probably unnecessary. Baevski, Auli, et al. (2020) self supervision enabling ASR at low cost

[5] It should be noted that ESPnet-EZ (Someki et al., 2024) provides a more modernized, python-focused interface similar to these generalist ML libraries while retaining some popular key features from Kaldi like recipes.

3 Related Work

In the following section, we will dive into some strategies used to address the problem areas outlined above: recent applications of ASR in MD, as well as some of the strategies adopted to overcoming the mismatch in objectives between ordinary phoneme recognition and MD which motivate the approaches taken in the current work:

1. ASR model ensembles, –and–

2. TTS-generated data augmentation

3.1 E2E ASR for MDD

Following the general trend in ASR and ML more broadly,

[20] Touch on phoneme extraction task and detail fine-tuning procedure Agrawal et al. (2023) smart weighter mechanism that selects model based on input audio

[21] MDD scoring

3.2 ASR Ensembling Strategies

(Bartelds et al., 2023; Kheir et al., 2023; Thai et al., 2019; D. Zhang et al., 2022, inter alia)

[22] how ensembling improves performance, different types As covered more extensively in the previous chapter, the state of modern ML has seen a proliferation of models trained on massive amounts of data tuned to perform well on a variety of targetet tasks, but can still struggle to perform satisfactorally outside their domain of expertise. One long-standing strategy for overcoming the limitations of any single classifier model in ASR as in ML more generally is through *ensemble* methods. In general terms, an ensemble refers to a set of classifiers whose outputs are combined in some way (Dietterich, 2000). This can be done in a number of different ways, Doing so effectively requires that these classifiers be similar in function but complementary, being diverse enough to cover each other's weaknesses (Hansen and Salamon, 1990)check source for their wording. Combining models was done originally by Bayesian averaging, i.e. weighing the output of the model by the posterior probability of the output, giving each model what amounts to a weighted vote in the output (Dietterich, 2000). Despite this mathematically principled origin, it proved to be prohibitively time consuming to calculate for complex problems, and so alternatives were introduced which were more computationally tractable. A wide range of ensembling techniques have emerged over the years, including boosting, bagging, stacking, majority voting and confidence voting, all of which have proven robust methods of improving performance on a variety of tasks (see l. Deng et al., 2012; Fiscus, 1997; Gitman et al., 2023; Maclin and Opitz, 1997, inter alia). For ASR, these approaches have yielded improvements over singular models with combinations of different model architectures (Arunkumar et al., 2022; L. Deng and Platt, 2014), and different training data such as for specific languages or dialects (Agrawal et al., 2023; Gitman et al., 2023, inter alia).

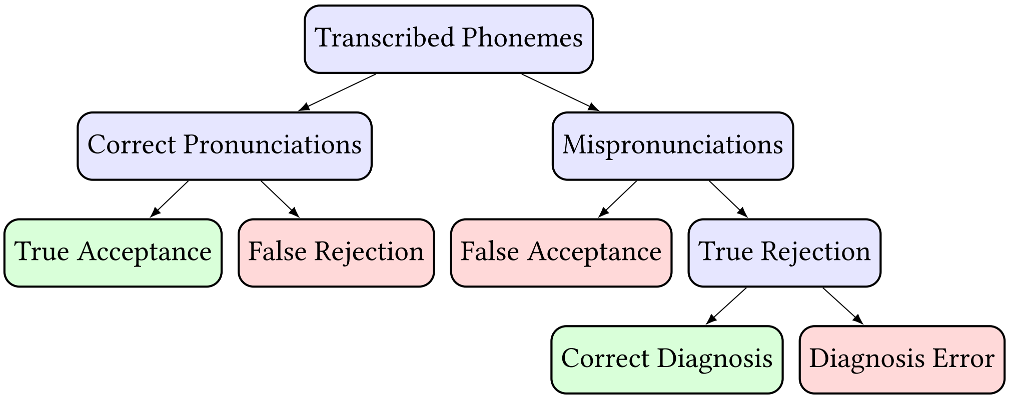

Figure 3.1: Evaluation hierarchy of phoneme level MD

[23] existing work in the area, gaps An early example of an equal weight, majority voting-based system in ASR is Recognizer Output Voting Error Reduction (ROVER) (Fiscus, 1997), whereby multiple systems are aligned with eachother which then vote on each aligned word slot, producing an aligned output string reached by model concensus. More dynamic approaches on the same theme have been explored as with Agrawal et al. (2023), whereby a weighting module is trained on transcriptions to select expert models so as to to minimize word error rate (WER), leading to an improvement over equal weighting. A number of confidence-based approaches have been taken towards language identification for both HMM and more modern E2E architectures Gitman et al. (2023), whereby only the outputs of the most confident model from the ensemble are used. One issue with deriving reliable confidence measures for modern E2E deep neural network (DNN)-based models lies in their tendency toward overconfidence in ouptput predictions (Wei et al., 2022), though some mitigation strategies have been introduced to derive more reliable confidence estimates from entropy calculations (Laptev and Ginsburg, 2023). Ensembles have been utilized in a number of ways for phone-level MD as well. Ananthakrishnan et al. (2011) investigated MD with a series of binary classifiers to detect mispronunciations typical of Swedish learners of varying L1 backgrounds, focusing on systematic phoneme confusion patterns over several utterances and exeplifying through two case studies how this approach could be used to motivate targeted interventions. Calık et al. (2023) also pursued MD for Arabic phonemes using a variety of model architectures and combination techniques i.e. bagging, boosting, stacking, and majority voting, finding majority voting the most performant. Combinations of models during training have also been pursued to ameliorate need for labeled training data through pseudo-labeling of unlabeled data (Yang et al., 2022), which connects to the motivation of our second experiment.

[24] Contribution of current work Although model ensembling has a long history in ML and ASR, confidence-based ensembled have not to our knowledge been leveraged toward the phone-level MD task. Despite the relative scarcity of ensembling applications to MD, the same limitations of data and model expertise clearly apply to L2-speech-specific phone classifiers as well. We have more readily available monolingual data to train expert models for Irish and English, but no labeled data available to train an equivalent expert model for speech more typical of Irish L2 speakers with an English L1. Models tuned to output phones for English would not be expected to reliably transcribe Irish-specific phones, and models tuned to Irish would similarly underperform for English phones. But given the possibility of combining these expert systems in an ensemble, could each constituent expert model's phone coverage be leveraged to complement the other's towards identifying mispronunciations in Irish L2 speech? If combined as a confidence-based ensemble of monolingual expert systems, we might for example expect an English model to more confidently predict English-like phones in an attempted Irish utterance with heavy interference from the speaker's L1, while a more native-like pronunciation for other segments without any L1 interference might be confidently transcribed by the Irish ASR. If we choose the most confidently predicted phone from each frame, we might be able to derive a combined string that more accurately reflects the segmental errors present in learner speech than is possible for each model on its own.

Hansen and Salamon (1990) Fiscus (1997) ROVER ensembling L. Deng and Platt (2014) Dietterich (2000) Jalalvand et al. (2018) novel ROVER approach with quality ranking at segment level Gitman et al. (2023) confidence-based ASR ensembles with selector block

3.3 Data Augmentation with Synthetic speech

Shen et al. (2018) TTS Tacotron (maybe find google tts paper) (Fazel et al., 2021)

[25] connect strategy to current work Punjabi et al. (2019) bootstrapping data with MT (important) Q. Xu et al. (2020) multiple iteration of pseudo-labeling on unlabeled data to overcome data scarcity Yang et al. (2022) pseudo-labeling to overcome scarcity (self-superviced learning)

Khare et al. (2021) transliteration as a bridge between orthographies Korzekwa et al. (2022) synthesis approaches to boosting asr Bartelds et al. (2023) TTS + ASR boosts ASR performance D. Zhang et al. (2022) phoneme paragraphing to generate mispronounced speech (important)

4 Methods & Materials

In the following chapter, we will elaborate the core experiments explored for the current work: one which explores model ensembling as described above as a solution to MD in the face of data scarcity; the other, which explores data augmentation via bootstrapping TTS input with grapheme-to-phoneme (G2P) model to create the data required to train an ASR model directly for MD. We will begin with detailing the features in common between these experiments, and end with the evaluation details shared between experiments.maybe setup links to refer to research questions or sections

4.1 General Experiment Setup

revisit connections to related work after that section is hammered out. At a high level, both of the MD pipelines explored follow the same general flow, starting with audio input of L2 speakers speaking Irish. The input waveform is then processed by one or several ASR models, configured to output strings of phonemes. These phoneme strings are then aligned to gold-standard transcriptions of the target string, allowing us to validate the performance of the ASR configuration through character error rate (CER). This string, together with derived posterior probabilities for frames corresponding to phonemes in the string, is used to compute GOP scores for the string's constituent phonemes, allowing us to gauge pronunciation quality.definitely don't feel like i understand

[26] base model description The model chosen for all experiments was the 300 million parameter version of Wav2vec2 XLS-R (Babu et al., 2021). This choice of architecture was motivated by the desire to leverage the aforementioned pre-training/fine-tuning paradigm on a model shown by Babu et al. (2021) to improve on previous work most notably for low- and mid-resource languages. Selection of the 300 million parameter version of this architecture was primarily motivated by desire to rapidly train models and test ideas, aiming for proof-of-concept over beating the state-of-the-art. These models were fine-tuned on phonetically annotated datasets, allowing us to obtain phoneme-level transcriptions for comparison.

All models were implemented using the Hugging Face *Transformers* library (Wolf et al., 2020). Training was managed with the Transformer *Trainer* class, configured via *TrainingArguments*. Training progress and metrics were logged to Weights & Biasesconfigure bibtex entry. Our implementation of Wav2Vec2 for phoneme output followed the publicly available tutorial by Vitouphy on Kaggle how to cite this appropriately. Models were instantiated with an attention dropout of 0.1, layer dropout of 0, feature projection dropout of 0, mask time probability of 0.75, mask time length of 10, mask feature probability of 0.25, and mask feature length of 64. create table with all this info.

Training for all ASR models models used identical parameters, with batch size of four, learning rate 3e-5, 2000 warmup steps, and 16-bit precision training to reduce memory usage and speed up training. CER was used in guiding early stopping, which had a patience of 6 epochs with a maximum of 30 epochs of training. create table with info. alternatively move this to the beginning to the other table if training is uniform (it should be)

4.2 Experiment 1: Monolingual Model Ensembling

[27] outline first experiment with model ensemble. Our first experiment starts with the ensembling intuition described abovemaybe reference ensembling section using twothree if using russian? expert systems: one Irish and one English. In contrast to previous works like L. Deng and Platt (2014)what was i saying here?. To yield posterior probability-like values from these model outputs, logit outputs of each model are normalized across the output vector for each frame. These normalized frame vectors are then compared to eachother with a heuristic selection mechanism: for every frame in the output vector, this block compares each component model output for that frame, and simply selects the model output with highest confidence. The combined output is then decoded, yielding a combined output string of phonemes with corresponding confidence values to be used at evaluation.

Guo et al. (2017) predicting probability estimates. how well calibrated are our models? Niehues and Pham (2019) similarity between training and test conditions for confidence (maybe more suitable as an extension) Papadopoulos et al. (2001) maximum likelihood, approximate bayesian, bootstrapping. (confidence estimation assessed by mean and st dev etc.) Wei et al. (2022) mitigating overconfidence with logit normalization during training. some background on softmax confidence scores

4.2.1 Data

[28] briefly summarize data sources To train our monolingual systems, we procured two monolingual corpora of read speech audio with phonetic transcriptions. Our English model was trained on the TIMIT corpus **garofolo1993timitempty citation** with the Irish model trained on audio data from the online Irish Dictionary and Language Library[1]. Krishenbaum (n.d.) use if ascii is used (it's not) Měchura (n.d.) introduction to grammar part of teanglann (not relevant? what did i want this for again?)

timit

The TIMIT corpus is a read speech corpus of American English speakers who spent their childhoods in one of eight major dialect areas of the United States. These areas are widely recognized with the exception of the Western dialect region, where boundaries are not confidently delineated, and the "army brat" group, consisting of speakers which frequently moved during their childhood due to the demands of highly mobile military service member parents, resulting in exposure to a variety of dialects(**garofoloDARPATIMITAcousticp** The full dataset consists of 6300 sentences from 630 speakers, so ten sentences per speaker. The duration of the data loaded from the Kaggle dataset linked to the TIMIT github dataset[2] lies at two hours and forty minutes. The TIMIT phonetic transcriptions were given in a revised version of ARPABET, a set of transcription codes developed by Advanced Research Projects Agency (ARPA)dig up reference for this.

teanglann

The Dictionary and Language Library is an online resource developed by Foras na Gaeilge[3] in conjunction with the New English-Irish Dictionary. Among the resources available are The Pronunciation Database, which contains recordings of individual

[1] https://www.teanglann.ie/en/

[2] https://github.com/philipperemy/timit

[3] Foras na Gaeilge is a group which promotes the Irish language, supports Irish-medium education, and advises public and private sector organizations, among its other functions. https://www.forasnagaeilge.ie/

words spoken by native speakers from three major dialects: Connacht, Ulster, and Munster. The Ulster recordings used for our monolingual Irish model were scraped by myself, adhering to the site's robots.txt file limit of one request per two seconds. The words scraped were drawn from an available G2P dictionarymaybe elaborate more where this comes from to ensure scraped audio had a corresponding canonical IPA transcriptions[4]. The resultant combined dataset had an audio duration of roughly five hours of audio. talk to jim about where he got the g2p dict I use for pronunciation information and as a scraping dict

Preprocessing

The audio from both Teanglann and TIMIT was loaded and resampled to 16kHz for compatability with Wav2vec2. For the phonetic trancriptions, TIMITs phonetic representations were translated to IPA using the phonecodes library, and maintaining one space between each phoneme or dipthongs[5]. The Irish data was transcribed through a lookup in the aforementioned G2P dictionary, which adhered to the same spacing rules as for TIMIT.

4.2.2 Experimental Setup

We split the datasets into train/dev/test splits with .8/.1/.1 ratios respectively, and then trained according to the above common . One critical limitation with the Irish data was that due to the acquisition method, we were left with no way verifiable way to identify speakers, so the different splits likely contains speaker overlap. We sought to mirror this limitation with the English data so as to not bias any particular language when ensembling them, and so the monolingual evaluation metrics are also likely overly optimistic.

4.3 Experiment 2: Synthetic Mispronunciation Data

The second experiment set out to capture mispronunciation by training a single ASR model whose base training data of more readily available, correctly pronounced native-speaker data was augmented with learner-like TTS-generated mispronounced versions of this base dataset. The input for generating this learner-like synthetic data is obtained through an iterative manual bootstrapping method whereby initial phonetic transcriptions of a set of words are adjusted to produce reasonable learner-like attempts to pronounce the canonical versions of each utterance via Speech Synthesis Markup Language (SSML) input to the TTS system. These manually adjusted transcriptions are then used to train a G2P model, from which we obtain a new round of phonetic transcriptions for the next set of words. The adjust-train-generate process is repeated until automatically generated transcriptions need only minimal if any adjustments.

4.3.1 Data

Canonical pronunciations were obtained from the Teanglann pronunciation dictionary, with corresponding canonical IPA transcriptions from the aforementioned Ulster Irish G2P dictionary: the same dataset as that used for training the Irish model in Experiment 4.2. The augmenting learner-like recordings were obtained from Google

[4] The dictionary of words to scrape was derived from an Ulster Irish G2P file generously provided by Jim O' Regan

[5] dipthongs were kept together as a single single unit to keep with standard practice

Cloud text-to-speechwhat to reference for this service? using the mention voice(s) used here voice(s) at a recording frequency of 16kHzdouble check the frequency. The input was given in SSML to enable direct use of IPA transcriptions to guide pronunciation. These correctly pronounced and corresponding mispronounced synthetic samples of each word composed the *positive* and *negative* utterance pairs of the training data. I will likely circle back to this distinction in the discussion section when suggesting an extension for contrastive training

4.3.2 Experimental Setup

Input was given as IPA strings in SSML format, The iterative improvement of TTS inputs was carried out by manual improvements of Montreal Forced Aligner (MFA) G2P (McAuliffe et al., 2017) in batches of 50 utterances at a time between retraining using default parametersis there anything interesting I can say about these models?. This process was repeated x times remember to complete this with the actual number of iterations, at which point the G2P-generated transcriptions ceased to improve meaningfully and required no changes to yield plausible learner speech. After training the above G2P model, recordings were produced from the TTS service for the remaining instances not used in G2P training. These recordings along with the IPA strings used to produce them were combined with aforementioned Teanglann recordings and their canonical transcriptions.

4.4 Evaluation

We evaluate phone-level discrimination of the ASR by first aligning model output strings to the *gold standard* annotated sequence with the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) using the Natural Language Toolkit (NLTK) library complete with whatever library I use for this.. From this alignment we can then compute phone error rate (PER) F1, Recall, all the other crap model off (Leung et al., 2019)

4.4.1 Metrics

For assessing the match of statistcal testing?

4.4.2 Data

[29] Common voice For evaluating the quality of our MD systems' sensitivity to plausible Irish L2 errors, we first needed Irish L2 speech. We begin with the Irish portion of the Common Voice datasetArdila et al. (2020), noted by Lonergan et al. (n.d.) as consisting nearly entirely of L2 speakers[6]. As in their work, this dataset will be used as the basis for testing the current experiment, as it fits the purpose of evaluating our system's effectiveness on L2 speech.this seems dumb, since wav2vec2 is trained on data before 2020. idlak maybe a better choice? The Common Voice corpus is a massive multilingual collection of transcribed speech designed for ASR which leverages crowd sourcing for data collection and validation to help alleviate the dirth of training data faced by most languages. why do I have three of the same paragraph

[30] annotation details To prepare the Common Voice dataset for use in phoneme-level MD, a small subset of the data was manually transcribed in IPA to capture pronunciation deviations from standard Ulster pronunciation. The process of capturing

[6] given its nature as a crowd-sourced dataset, some data has been added since that date, but its character as a largely if not entirely L2 speaker-dominated set of Irish speech should remain the case

these deviations consisted of the following steps: first, an IPA representation of the transcription was generated from a lookup in the Ulster G2P dictionary to map the orthographic transcriptions to their canonical IPA representations; recordings were then manually assessed, comparing these generated IPA representations to the audio and noting where the phonemes deviated from the canonical pronunciations and what their realization was assessed to be. This process was carried out by myself only, a notable limitation we will return to later. The annotations were carried out in Label Studio using some preliminary annotations generated by a lookup in the aforementioned Ulster G2P dictionary with a MFA G2P model trained on this dictionary to serve as a fallback when the lookup returned no match.ask about where the g2p came from, what to reference

Conneau et al. (2022) Parallel ASR dataset. move to discussion. Deichler et al. (2024) multimodal conversational dataset with cospeech gestures. move to discussion Qian et al. (2022) ASR for irish: uses Mozilla common voice J. Zhang et al. (2021) open source speech corpus speech ocean for pronunciation assessment Zhao et al. (2018) L2 arctic dataset Lee et al. (n.d.) wikipron (not used but illustrative as a reference)

[31] metrics CER, F1, etc. Detail CTC loss used in training CTC loss Graves et al. (n.d.) also details some peakiness Kürzinger et al. (2020) ctc and dataset bootstrapping

4.4.3 Skyline Comparison

To see how our approaches compare to more idealize skyline conditions we compare performance of the above experiments to the average performance of four models trained on English L2 speakers with a commonly used dataset for MD research: L2ARCTIC (Zhao et al., 2018). L2-ARCTIC is a read-speech corpus of non-native English, originally released with speakers of five different L1s: Hindi, Korean, Mandarin, Spanish, and Arabic, though Vietnamese has since been added. At present, each language has four speakers: two male and two female, from whom over an hour of speech recordings were taken along with word- and phoneme-level transcription, and selected manual annotations for each recording. To train our skyline models, we confined ourself to one speaker L1 group, Spanish. This was to allow our models to learn a single L1 interference profile relatively close to English while minimizing heterogeneity in L1 interference Still not sure what I think about this sentence. The phoneme-level annotations were presented in ARPABET, from which we could obtain IPA phonetic transcriptions to train models using the same training strategy outlined above. Due to the limited speaker pool of four for each language, and within that, two for each gender, any one model trained with a held out speaker for evaluation ran the risk of overfitting to a specific gender profile. To combat this, we ran 4-fold leave-one-speaker-out (LOSO) cross-validation (CV), using the average performance to represent our skyline performance ceiling.

5 Results

[32] detail the results as it pertains to the evaluation framework: CER, MDD metrics, etc

6 Discussion

[33] Summarize the main findings from results and how it relates to the research questions

[34] why confidence might need adjusting? (Laptev and Ginsburg, 2023)

[35] Detail possible applications to Language learning, the role of results in augmenting self-directed language learning Hardison (2005) effect of multimodal input on speech identification

[36] Detail possible (or actual if time allows) Furhat application **deichler2024mmempty citation** multimedia data

[37] continue furhat explanation

[38] Argue for accurate articulator representation in digital agents as supportive pillar in pronunciation feedback. **Li2011ThePAempty citation** Animated articulators Rosenblum (2008) speech perception as multimodal phenomenom

7 Conclusions

[39] reiterate how results connect with research questions, set main conclusions in context of impact to research and society Zeyer et al. (2021) peaky ctc, ambiguous phone boundaries

[40] Extensions from previous research (yang2022) Liu et al. (2023) wav2vec which does adversarial training to improve discrimination. Peng et al. (2023) contrastive loss optimization Lonergan et al. (2024) alternative architectures for low resource asr Neri et al. (2008) pronunciation training for children, lead to improvements comparable to traditional training Prabhavalkar et al. (2017) alternatives to ctc

[41] Possibilities for future work Gong et al. (2022) assessment targeting more than one aspect of speech (prosody, word-level stress) Mortensen et al. (n.d.) PanPhon mapping from ipa to articulatory features Rouditchenko et al. (2023) language family in pretraining predictive of how models compare. need for resources for smaller families Sjons (2022) focus on child directed speech, since it is well-suited for word segmentation?

8 Ethical Considerations

[42] General Ethical considerations for current work and possible extensions
[43] Ethical considerations for Minority languages

9 AI Tools

[44] Describe use of AI tools and how its use benefited me.

Bibliography

Agrawal, Aakriti, Milind Rao, Anit Kumar Sahu, Gopinath Chennupati, and Andreas Stolcke (2023). “Learning When to Trust Which Teacher for Weakly Supervised ASR”. In: *INTERSPEECH 2023*. INTERSPEECH 2023. ISCA, Aug. 20, 2023, pp. 381–385. DOI: 10.21437/Interspeech.2023-2205. URL: https://www.isca-archive.org/interspeech_2023/agrawal23_interspeech.html (visited on 2025-06-16).

Ananthakrishnan, Gopal, Preben Wik, Olov Engwall, and Sherif Abdou (2011). “Using an Ensemble of Classifiers for Mispronunciation Feedback”. In: *Speech and Language Technology in Education (SLaTE 2011)*. Speech and Language Technology in Education (SLaTE 2011). ISCA, Aug. 24, 2011, pp. 49–52. DOI: 10.21437/SLaTE.2011-13. URL: https://www.isca-archive.org/slate_2011/ananthakrishnan11_slate.html (visited on 2025-09-17).

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber (2020). *Common Voice: A Massively-Multilingual Speech Corpus*. Mar. 5, 2020. DOI: 10.48550/arXiv.1912.06670. arXiv: 1912.06670 [cs]. URL: http://arxiv.org/abs/1912.06670 (visited on 2025-04-04). Pre-published.

Arunkumar, A., Vrunda N. Sukhadia, and S. Umesh (2022). “Investigation of Ensemble Features of Self-Supervised Pretrained Models for Automatic Speech Recognition”. In: *Interspeech 2022*. Sept. 18, 2022, pp. 5145–5149. DOI: 10.21437/Interspeech.2022-11376. arXiv: 2206.05518 [cs]. URL: http://arxiv.org/abs/2206.05518 (visited on 2025-08-29).

Atlas of the World’s Languages in Danger (2010). Paris: UNESCO Publishing.

Babu, Arun, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli (2021). *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. Dec. 16, 2021. DOI: 10.48550/arXiv.2111.09296. arXiv: 2111.09296 [cs]. URL: http://arxiv.org/abs/2111.09296 (visited on 2025-09-07). Pre-published.

Baevski, Alexei, Michael Auli, and Abdelrahman Mohamed (2020). *Effectiveness of Self-Supervised Pre-Training for Speech Recognition*. May 18, 2020. DOI: 10.48550/arXiv.1911.03912. arXiv: 1911.03912 [cs]. URL: http://arxiv.org/abs/1911.03912 (visited on 2025-04-04). Pre-published.

Baevski, Alexei, Steffen Schneider, and Michael Auli (2020). *Vq-Wav2vec: Self-Supervised Learning of Discrete Speech Representations*. Feb. 16, 2020. DOI: 10.48550/arXiv.1910.05453. arXiv: 1910.05453 [cs]. URL: http://arxiv.org/abs/1910.05453 (visited on 2025-04-04). Pre-published.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). “Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 12449–12460. URL: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html (visited on 2025-09-09).

Bajorek, Joan Palmiter (2017). “L2 Pronunciation in CALL: The Unrealized Potential of Rosetta Stone, Duolingo, Babbel, and Mango Languages”. *Issues and Trends in Educational Technology* 5.1, pp. 24–51.

Bartelds, Martijn, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling (2023). *Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation.* May 19, 2023. DOI: 10.48550/arXiv.2305.10951. arXiv: 2305.10951 [cs]. URL: http://arxiv.org/abs/2305.10951 (visited on 2025-07-15). Pre-published.

Bender, Emily M. (2011). "On Achieving and Evaluating Language-Independence in NLP". *Linguistic Issues in Language Technology* 6 (Oct. 1, 2011). ISSN: 1945-3604. DOI: 10.33011/lilt.v6i.1239. URL: https://journals.colorado.edu/index.php/lilt/article/view/1239 (visited on 2025-04-04).

Besacier, Laurent, Etienne Barnard, Alexey Karpov, and Tanja Schultz (2014). "Automatic Speech Recognition for Under-Resourced Languages: A Survey". *Speech Communication* 56 (Jan. 2014), pp. 85–100. ISSN: 0167-6393. DOI: 10.1016/j.specom.2013.07.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639313000988 (visited on 2025-07-15).

Broin, Brian Ó (2014). "New Urban Irish: Pidgin, Creole, or Bona Fide Dialect? The Phonetics and Morphology of City and Gaeltacht Speakers Systematically Compared".

Calık, Sükrü Selim, Ayhan Kucukmanisa, and Zeynep Hilal Kilimci (2023). "An Ensemble-Based Framework for Mispronunciation Detection of Arabic Phonemes". *Applied Acoustics* 212 (Sept. 2023), p. 109593. ISSN: 0003682X. DOI: 10.1016/j.apacoust.2023.109593. URL: https://linkinghub.elsevier.com/retrieve/pii/S0003682X23003912 (visited on 2025-09-17).

Chasaide, Ailbhe, Neasa Ní Chiaráin, Christoph Wendler, Harald Berthelsen, Andy Murphy, and Christer Gobl (2017). *The ABAIR Initiative: Bringing Spoken Irish into the Digital Space.* Aug. 20, 2017. DOI: 10.21437/Interspeech.2017-1407.

Conneau, Alexis, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli (2020). *Unsupervised Cross-lingual Representation Learning for Speech Recognition.* Dec. 15, 2020. DOI: 10.48550/arXiv.2006.13979. arXiv: 2006.13979 [cs]. URL: http://arxiv.org/abs/2006.13979 (visited on 2025-04-04). Pre-published.

Conneau, Alexis, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna (2022). *FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech.* May 25, 2022. DOI: 10.48550/arXiv.2205.12446. arXiv: 2205.12446 [cs]. URL: http://arxiv.org/abs/2205.12446 (visited on 2025-04-04). Pre-published.

Deichler, Anna, Jim O'Regan, and Jonas Beskow (2024). *MM-Conv: A Multi-modal Conversational Dataset for Virtual Humans.* Sept. 30, 2024. DOI: 10.48550/arXiv.2410.00253. arXiv: 2410.00253 [cs]. URL: http://arxiv.org/abs/2410.00253 (visited on 2025-04-04). Pre-published.

Deng, Li and John C. Platt (2014). "Ensemble Deep Learning for Speech Recognition". In: *Proc. Interspeech.*

Deng, li, Dong Yu, and John Platt (2012). "Scalable Stacking and Learning for Building Deep Architectures". *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on* (Mar. 1, 2012). DOI: 10.1109/ICASSP.2012.6288333.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* NAACL-HLT 2019. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/ (visited on 2025-08-28).

Dietterich, Thomas G (2000). “Ensemble Methods in Machine Learning”. In: International Workshop on Multiple Classifier Systems. Springer, pp. 1–15. URL: https://www2.cs.uh.edu/~ceick/7362/T5-3.pdf (visited on 2025-09-06).

Dryer, Matthew, Martin Haspelmath, Matthew Dryer, and Martin Haspelmath (2024). *The World Atlas of Language Structures Online*. Version v2020.4. Zenodo, Oct. 18, 2024. DOI: 10.5281/ZENODO.13950591. URL: https://zenodo.org/doi/10.5281/zenodo.13950591 (visited on 2025-07-08).

Engstrand, Olle (2004). *Fonetikens Grunder*. Lund: Studentlitteratur. 355 pp. ISBN: 978-91-44-04238-1.

Fazel, Amin, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo (2021). *SynthASR: Unlocking Synthetic Data for Speech Recognition*. June 14, 2021. DOI: 10.48550/arXiv.2106.07803. arXiv: 2106.07803 [cs]. URL: http://arxiv.org/abs/2106.07803 (visited on 2025-09-13). Pre-published.

Fiscus, Jonathan G (1997). “A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)”. In: *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. IEEE, pp. 347–354.

Gabriele, Jennifer C (n.d.). “English Influence on L2 Speakers’ Production of Palatalization and Velarization” ().

Gitman, Igor, Vitaly Lavrukhin, Aleksandr Laptev, and Boris Ginsburg (2023). “Confidence-Based Ensembles of End-to-End Speech Recognition Models”. In: *INTERSPEECH 2023*. Aug. 20, 2023, pp. 1414–1418. DOI: 10.21437/Interspeech.2023-1281. arXiv: 2306.15824 [eess]. URL: http://arxiv.org/abs/2306.15824 (visited on 2025-04-04).

Gong, Yuan, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass (2022). “Transformer-Based Multi-Aspect Multi-Granularity Non-Native English Speaker Pronunciation Assessment”. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 23, 2022, pp. 7262–7266. DOI: 10.1109/ICASSP43922.2022.9746743. arXiv: 2205.03432 [cs]. URL: http://arxiv.org/abs/2205.03432 (visited on 2025-04-04).

Graves, Alex, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber (n.d.). “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks” ().

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger (2017). *On Calibration of Modern Neural Networks*. Aug. 3, 2017. DOI: 10.48550/arXiv.1706.04599. arXiv: 1706.04599 [cs]. URL: http://arxiv.org/abs/1706.04599 (visited on 2025-06-16). Pre-published.

Hansen, L.K. and P. Salamon (1990). “Neural Network Ensembles”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.10 (Oct. 1990), pp. 993–1001. ISSN: 01628828. DOI: 10.1109/34.58871. URL: http://ieeexplore.ieee.org/document/58871/ (visited on 2025-09-14).

Hardison, Debra M. (2005). “Second-Language Spoken Word Identification: Effects of Perceptual Training, Visual Cues, and Phonetic Environment”. *Applied Psycholinguistics* 26.4 (Oct. 2005), pp. 579–596. ISSN: 0142-7164, 1469-1817. DOI: 10.1017/S0142716405050319. URL: https://www.cambridge.org/core/product/identifier/S0142716405050319/type/journal_article (visited on 2025-04-04).

Hedderich, Michael A., Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow (2021). *A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios*. Apr. 9, 2021. DOI: 10.48550/arXiv.2010.12309. arXiv: 2010.12309 [cs]. URL: http://arxiv.org/abs/2010.12309 (visited on 2025-04-04). Pre-published.

Jalalvand, Shahab, Matteo Negri, Daniele Falavigna, Marco Matassoni, and Marco Turchi (2018). “Automatic Quality Estimation for ASR System Combination”. *Computer Speech & Language* 47 (Jan. 2018), pp. 214–239. ISSN: 08852308. DOI: 10.1016/j.

csl.2017.06.003. arXiv: 1706.07238 [cs]. URL: http://arxiv.org/abs/1706.07238 (visited on 2025-06-16).

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2021). *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*. Jan. 27, 2021. DOI: 10.48550/arXiv.2004.09095. arXiv: 2004.09095 [cs]. URL: http://arxiv.org/abs/2004.09095 (visited on 2025-04-04). Pre-published.

Jurafsky, Dan and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd. URL: https://web.stanford.edu/~jurafsky/slp3/.

Khare, Shreya, Ashish Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj (2021). "Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration". In: *Interspeech 2021*. Interspeech 2021. ISCA, Aug. 30, 2021, pp. 1529–1533. DOI: 10.21437/Interspeech.2021-2062. URL: https://www.isca-archive.org/interspeech_2021/khare21_interspeech.html (visited on 2025-09-17).

Kheir, Yassine El, Ahmed Ali, and Shammur Absar Chowdhury (2023). *Automatic Pronunciation Assessment – A Review*. Oct. 21, 2023. DOI: 10.48550/arXiv.2310.13974. arXiv: 2310.13974 [cs]. URL: http://arxiv.org/abs/2310.13974 (visited on 2025-04-04). Pre-published.

Korzekwa, Daniel, Jaime Lorenzo-Trueba, Thomas Drugman, and Bozena Kostek (2022). "Computer-Assisted Pronunciation Training—Speech Synthesis Is Almost All You Need". *Speech Communication* 142 (July 2022), pp. 22–33. ISSN: 0167-6393. DOI: 10.1016/j.specom.2022.06.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639322000863 (visited on 2025-07-18).

Krashen, Stephen D. (1984). *Principles and Practice in Second Language Acquisition*. Reprinted. Language Teaching Methodology Series. Oxford: Pergamon Press. 202 pp. ISBN: 978-0-08-028628-0.

Krishenbaum, Evan (n.d.). "Representing IPA Phonetics in ASCII" ().

Kürzinger, Ludwig, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll (2020). "CTC-Segmentation of Large Corpora for German End-to-end Speech Recognition". In: vol. 12335, pp. 267–278. DOI: 10.1007/978-3-030-60276-5_27. arXiv: 2007.09127 [eess]. URL: http://arxiv.org/abs/2007.09127 (visited on 2025-04-04).

Laptev, Aleksandr and Boris Ginsburg (2023). "Fast Entropy-Based Methods of Word-Level Confidence Estimation for End-To-End Automatic Speech Recognition". In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. Jan. 9, 2023, pp. 152–159. DOI: 10.1109/SLT54892.2023.10022960. arXiv: 2212.08703 [eess]. URL: http://arxiv.org/abs/2212.08703 (visited on 2025-09-14).

Lee, Jackson L, Lucas F E Ashby, M Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D McCarthy, and Kyle Gorman (n.d.). "Massively Multilingual Pronunciation Mining with WikiPron" ().

Leung, Wai-Kim, Xunying Liu, and Helen Meng (2019). "CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, May 2019, pp. 8132–8136. ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8682654. URL: https://ieeexplore.ieee.org/document/8682654/ (visited on 2025-09-12).

Liu, Alexander H., Wei-Ning Hsu, Michael Auli, and Alexei Baevski (2023). "Towards End-to-End Unsupervised Speech Recognition". In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. 2022 IEEE Spoken Language Technology Workshop (SLT). Doha, Qatar: IEEE, Jan. 9, 2023, pp. 221–228. ISBN: 979-8-3503-9690-4. DOI: 10.1109/

SLT54892.2023.10023187. URL: https://ieeexplore.ieee.org/document/10023187/ (visited on 2025-04-04).

Lo, Tien-Hong, Shi-Yan Weng, Hsiu-Jui Chang, and Berlin Chen (2020). *An Effective End-to-End Modeling Approach for Mispronunciation Detection.* May 18, 2020. DOI: 10.48550/arXiv.2005.08440. arXiv: 2005.08440 [eess]. URL: http://arxiv.org/abs/2005.08440 (visited on 2025-08-30). Pre-published.

Lonergan, Liam, Mengjie Qian, Harald Berthelsen, Andy Murphy, Christoph Wendler, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide (n.d.). "Automatic Speech Recognition for Irish: The ABAIR-ÉIST System" ().

Lonergan, Liam, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide (2024). *Low-Resource Speech Recognition and Dialect Identification of Irish in a Multi-Task Framework.* May 2, 2024. DOI: 10.48550/arXiv.2405.01293. arXiv: 2405.01293 [cs]. URL: http://arxiv.org/abs/2405.01293 (visited on 2025-04-04). Pre-published.

Maclin, Richard and David Opitz (1997). "An Empirical Evaluation of Bagging and Boosting". *AAAI/IAAI* 1997, pp. 546–551. URL: https://d1wqtxts1xzle7.cloudfront.net/78173773/An_Empirical_Evaluation_of_Bagging_and_B20220105-9764-16yot8b.pdf?1738464338=&response-content-disposition=inline%3B+filename%3DAn_empirical_evaluation_of_bagging_and_b.pdf&Expires=1758100511&Signature=ULr42EX~oPCHa6soKPNoavlVoaJeaV5Pq96lNHezXzPN3CaB2CYdPJi1atYw3dXBQFo4naJPYSQicHHyAQrqLNgXwSF8xAx6EFU4g~CBnmOa5vnYQeQJTrcJlOvLUwbAHWoiZf8T8lNgeP-c~aImdq1OYrzWsEe2HQ6e-q77emb2VM9p~xsFxMKAn~5RU2w6d~x3iQ4n-YVaAeUbPdAtWFlysMTlSow56C4xK4KupsOSUA3qlXeq9PTBaWskjFVCcSAlfJpmo6hDOLrZtwO3O821qdsrh1HsXxoTTWa2Qbh1~Ra-ZHwXQt-02~XKDva8-HeQxOX7x~6FrBsf7HBo6A__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA (visited on 2025-09-17).

Magueresse, Alexandre, Vincent Carles, and Evan Heetderks (2020). *Low-Resource Languages: A Review of Past Work and Future Challenges.* June 12, 2020. DOI: 10.48550/arXiv.2006.07264. arXiv: 2006.07264 [cs]. URL: http://arxiv.org/abs/2006.07264 (visited on 2025-04-04). Pre-published.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi". In: *Interspeech 2017.* Interspeech 2017. ISCA, Aug. 20, 2017, pp. 498–502. DOI: 10.21437/Interspeech.2017-1386. URL: https://www.isca-archive.org/interspeech_2017/mcauliffe17_interspeech.html (visited on 2025-09-08).

Měchura, Michal Boleslav (n.d.). "Introduction to Gramadán and the Irish National Morphology Database" ().

Mortensen, David R, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin (n.d.). "PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors" ().

Needleman, Saul B and Christian D Wunsch (1970). "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins". *Journal of molecular biology* 48.3, pp. 443–453.

Neri, Ambra, Ornella Mich, Matteo Gerosa, and Diego Giuliani (2008). "The Effectiveness of Computer Assisted Pronunciation Training for Foreign Language Learning by Children". *Computer Assisted Language Learning* 21.5 (Dec. 2008), pp. 393–408. ISSN: 0958-8221, 1744-3210. DOI: 10.1080/09588220802447651. URL: https://www.tandfonline.com/doi/full/10.1080/09588220802447651 (visited on 2025-04-04).

Ní Chasaide, Ailbhe, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, and Andrew Murphy (2015). "SPEECH TECHNOLOGY AS DOCUMENTATION FOR ENDANGERED LANGUAGE PRESERVATION: THE CASE OF IRISH".

Ní Chasaide, Ailbhe, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andrew Murphy, Emily Barnes, and Christer Gobl (2019). “Can We Defuse the Digital Timebomb? Linguistics, Speech Technology and the Irish Language Community”. In: *Proceedings of the Language Technologies for All (LT4All)*. Proceedings of the Language Technologies for All (LT4All). European Language Resources Association (ELRA), pp. 177–181. DOI: 10.21437/SpeechProsody.2016-73. URL: https://lt4all.elra.info/media/papers/O8/97.pdf (visited on 2025-04-04).

Niehues, Jan and Ngoc-Quan Pham (2019). *Modeling Confidence in Sequence-to-Sequence Models*. Oct. 4, 2019. DOI: 10.48550/arXiv.1910.01859. arXiv: 1910.01859 [cs]. URL: http://arxiv.org/abs/1910.01859 (visited on 2025-04-04). Pre-published.

Papadopoulos, G., P.J. Edwards, and A.F. Murray (2001). “Confidence Estimation Methods for Neural Networks: A Practical Comparison”. *IEEE Transactions on Neural Networks* 12.6 (Nov. 2001), pp. 1278–1287. ISSN: 10459227. DOI: 10.1109/72.963764. URL: http://ieeexplore.ieee.org/document/963764/ (visited on 2025-04-04).

Parikh, Aditya Kamlesh, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik (2025). *Evaluating Logit-Based GOP Scores for Mispronunciation Detection*. July 8, 2025. DOI: 10.48550/arXiv.2506.12067. arXiv: 2506.12067 [eess]. URL: http://arxiv.org/abs/2506.12067 (visited on 2025-08-31). Pre-published.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (visited on 2025-09-08).

Peng, Linkai, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan (2021). “A Study on Fine-Tuning Wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis”. In: *Interspeech 2021*. Interspeech 2021. ISCA, Aug. 30, 2021, pp. 4448–4452. DOI: 10.21437/Interspeech.2021-1344. URL: https://www.isca-archive.org/interspeech_2021/peng21e_interspeech.html (visited on 2025-04-04).

Peng, Linkai, Yingming Gao, Rian Bao, Ya Li, and Jinsong Zhang (2023). “End-to-End Mispronunciation Detection and Diagnosis Using Transfer Learning”. *Applied Sciences* 13.11 (June 2, 2023), p. 6793. ISSN: 2076-3417. DOI: 10.3390/app13116793. URL: https://www.mdpi.com/2076-3417/13/11/6793 (visited on 2025-06-16).

Peng, Linkai, Yingming Gao, Binghuai Lin, Dengfeng Ke, Yanlu Xie, and Jinsong Zhang (2022). *Text-Aware End-to-end Mispronunciation Detection and Diagnosis*. June 15, 2022. DOI: 10.48550/arXiv.2206.07289. arXiv: 2206.07289 [cs]. URL: http://arxiv.org/abs/2206.07289 (visited on 2025-06-16). Pre-published.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukaˇs Burget, Ondˇrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlıˇcek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely (n.d.). “The Kaldi Speech Recognition Toolkit” ().

Prabhavalkar, Rohit, Tara N. Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, and Anjuli Kannan (2017). *Minimum Word Error Rate Training for Attention-based Sequence-to-Sequence Models*. Dec. 5, 2017. DOI: 10.48550/arXiv.1712.01818. arXiv: 1712.01818 [cs]. URL: http://arxiv.org/abs/1712.01818 (visited on 2025-04-04). Pre-published.

Punjabi, Surabhi, Harish Arsikere, and Sri Garimella (2019). “Language Model Bootstrapping Using Neural Machine Translation for Conversational Speech Recognition”. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

SG, Singapore: IEEE, Dec. 2019, pp. 487–493. ISBN: 978-1-7281-0306-8. DOI: 10.1109/ASRU46091.2019.9003982. URL: https://ieeexplore.ieee.org/document/9003982/ (visited on 2025-06-16).

Qian, Mengjie, Harald Berthelsen, Liam Lonergan, Andy Murphy, Claire O'Neill, Neasa Ni Chiarain, Christer Gobl, and Ailbhe Ni Chasaide (2022). "Automatic Speech Recognition for Irish: Testing Lexicons and Language Models". In: *2022 33rd Irish Signals and Systems Conference (ISSC)*. 2022 33rd Irish Signals and Systems Conference (ISSC). Cork, Ireland: IEEE, June 9, 2022, pp. 1–6. ISBN: 978-1-6654-5227-4. DOI: 10.1109/ISSC55427.2022.9826201. URL: https://ieeexplore.ieee.org/document/9826201/ (visited on 2025-04-04).

Ranathunga, Surangika, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur (2021). *Neural Machine Translation for Low-Resource Languages: A Survey*. June 29, 2021. DOI: 10.48550/arXiv.2106.15115. arXiv: 2106.15115 [cs]. URL: http://arxiv.org/abs/2106.15115 (visited on 2025-04-04). Pre-published.

Rogerson-Revell, Pamela M (2021). "Computer-Assisted Pronunciation Training (CAPT): Current Issues and Future Directions". *RELC Journal* 52.1 (Apr. 2021), pp. 189–205. ISSN: 0033-6882, 1745-526X. DOI: 10.1177/0033688220977406. URL: https://journals.sagepub.com/doi/10.1177/0033688220977406 (visited on 2025-04-04).

Rosenblum, Lawrence D. (2008). "Speech Perception as a Multimodal Phenomenon". *Current Directions in Psychological Science* 17.6 (Dec. 2008), pp. 405–409. ISSN: 0963-7214, 1467-8721. DOI: 10.1111/j.1467-8721.2008.00615.x. URL: https://journals.sagepub.com/doi/10.1111/j.1467-8721.2008.00615.x (visited on 2025-04-04).

Rouditchenko, Andrew, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass (2023). "Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages". In: *INTERSPEECH 2023*. INTERSPEECH 2023. ISCA, Aug. 20, 2023, pp. 2268–2272. DOI: 10.21437/Interspeech.2023-1061. URL: https://www.isca-archive.org/interspeech_2023/rouditchenko23_interspeech.html (visited on 2025-04-04).

Shahin, Mostafa and Beena Ahmed (2024). "Phonological-Level Mispronunciation Detection and Diagnosis". In: *Interspeech 2024*. Interspeech 2024. ISCA, Sept. 1, 2024, pp. 307–311. DOI: 10.21437/Interspeech.2024-2217. URL: https://www.isca-archive.org/interspeech_2024/shahin24_interspeech.html (visited on 2025-04-04).

Shen, Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu (2018). *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. Feb. 16, 2018. DOI: 10.48550/arXiv.1712.05884. arXiv: 1712.05884 [cs]. URL: http://arxiv.org/abs/1712.05884 (visited on 2025-04-04). Pre-published.

Sjons, Johan (2022). "Articulation Rate and Surprisal in Swedish Child-Directed Speech". PhD thesis. Stockholm University.

Snesareva, Marina (2016). "Palatalization in Dublin Irish: The Extent of Phonetic Interference". *Procedia - Social and Behavioral Sciences* 236 (Dec. 2016), pp. 213–218. ISSN: 18770428. DOI: 10.1016/j.sbspro.2016.12.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877042816316421 (visited on 2025-04-04).

Someki, Masao, Kwanghee Choi, Siddhant Arora, William Chen, Samuele Cornell, Jionghao Han, Yifan Peng, Jiatong Shi, Vaibhav Srivastav, and Shinji Watanabe (2024). *ESPnet-EZ: Python-only ESPnet for Easy Fine-tuning and Integration*. Sept. 14, 2024. DOI: 10.48550/arXiv.2409.09506. arXiv: 2409.09506 [cs]. URL: http://arxiv.org/abs/2409.09506 (visited on 2025-04-04). Pre-published.

Spolsky, Bernard and Francis M Hult (2008). “The Handbook of Educational Linguistics”. *Wiley Online Library*.

Stanley, Theban and Kadri Hacioglu (2012). “Improving L1-specific Phonological Error Diagnosis in Computer Assisted Pronunciation Training”. In: *Interspeech 2012*. Interspeech 2012. ISCA, Sept. 9, 2012, pp. 827–830. DOI: 10.21437/Interspeech.2012-251. URL: https://www.isca-archive.org/interspeech_2012/stanley12_interspeech.html (visited on 2025-04-04).

Stenson, Nancy (2020). *Modern Irish: A Comprehensive Grammar*. Routledge Comprehensive Grammars. London ; New York: Routledge, Taylor & Francis. 304 pp. ISBN: 978-1-138-23652-3 978-1-138-23651-6.

Thai, Bao, Robert Jimerson, Dominic Arcoraci, Emily Prud’hommeaux, and Raymond Ptucha (2019). “Synthetic Data Augmentation for Improving Low-Resource ASR”. In: *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*. 2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW). Rochester, NY, USA: IEEE, Oct. 2019, pp. 1–9. ISBN: 978-1-7281-4352-1. DOI: 10.1109/WNYIPW.2019.8923082. URL: https://ieeexplore.ieee.org/document/8923082/ (visited on 2025-08-28).

Watanabe, Shinji, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai (2018). *ESPnet: End-to-End Speech Processing Toolkit*. Mar. 30, 2018. DOI: 10.48550/arXiv.1804.00015. arXiv: 1804.00015 [cs]. URL: http://arxiv.org/abs/1804.00015 (visited on 2025-09-03). Pre-published.

Wei, Hongxin, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li (2022). *Mitigating Neural Network Overconfidence with Logit Normalization*. June 24, 2022. DOI: 10.48550/arXiv.2205.09310. arXiv: 2205.09310 [cs]. URL: http://arxiv.org/abs/2205.09310 (visited on 2025-06-16). Pre-published.

Witt, Silke and Steve Young (2014). “Computer-Assisted Pronunciation Teaching Based on Automatic Speech Recognition”. In: *Language Teaching and Language Technology*. Routledge, pp. 25–35.

Witt, Silke M (n.d.). “Automatic Error Detection in Pronunciation Training: Where We Are and Where We Need to Go” ().

Witt, Silke Maren (2000). “Use of Speech Recognition in Computer-Assisted Language Learning.” PhD thesis.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (2020). “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6/ (visited on 2025-09-08).

Wu, Minglin, Kun Li, Wai-Kim Leung, and Helen Meng (2021). “Transformer Based End-to-End Mispronunciation Detection and Diagnosis”. In: *Interspeech 2021*. Interspeech 2021. ISCA, Aug. 30, 2021, pp. 3954–3958. DOI: 10.21437/Interspeech.2021-1467. URL: https://www.isca-archive.org/interspeech_2021/wu21h_interspeech.html (visited on 2025-04-04).

Xu, Qiantong, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert (2020). “Iterative Pseudo-Labeling for Speech Recognition”. In: *Interspeech 2020*. Interspeech 2020. ISCA, Oct. 25, 2020, pp. 1006–1010. DOI: 10.

21437/Interspeech.2020-1800. URL: https://www.isca-archive.org/interspeech_2020/xu20b_interspeech.html (visited on 2025-04-04).

Xu, Xiaoshuo, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma (2021). "Explore Wav2vec 2.0 for Mispronunciation Detection". In: *Interspeech 2021*. Interspeech 2021. ISCA, Aug. 30, 2021, pp. 4428–4432. DOI: 10.21437/Interspeech.2021-777. URL: https://www.isca-archive.org/interspeech_2021/xu21k_interspeech.html (visited on 2025-04-04).

Yang, Mu, Kevin Hirschi, Stephen Daniel Looney, Okim Kang, and John H.L. Hansen (2022). "Improving Mispronunciation Detection with Wav2vec2-based Momentum Pseudo-Labeling for Accentedness and Intelligibility Assessment". In: *Interspeech 2022*. Interspeech 2022. ISCA, Sept. 18, 2022, pp. 4481–4485. DOI: 10.21437/Interspeech.2022-11039. URL: https://www.isca-archive.org/interspeech_2022/yang22v_interspeech.html (visited on 2025-04-04).

Yu, Dong and Li Deng (2015). *Automatic Speech Recognition: A Deep Learning Approach*. Signals and Communication Technology. London: Springer London. ISBN: 978-1-4471-5778-6 978-1-4471-5779-3. DOI: 10.1007/978-1-4471-5779-3. URL: https://link.springer.com/10.1007/978-1-4471-5779-3 (visited on 2025-08-31).

Zeyer, Albert, Ralf Schlüter, and Hermann Ney (2021). *Why Does CTC Result in Peaky Behavior?* June 3, 2021. DOI: 10.48550/arXiv.2105.14849. arXiv: 2105.14849 [cs]. URL: http://arxiv.org/abs/2105.14849 (visited on 2025-06-16). Pre-published.

Zhang, Daniel, Ashwinkumar Ganesan, Sarah Campbell, and Daniel Korzekwa (2022). "L2-GEN: A Neural Phoneme Paraphrasing Approach to L2 Speech Synthesis for Mispronunciation Diagnosis". In: *Interspeech 2022*. Interspeech 2022. ISCA, Sept. 18, 2022, pp. 4317–4321. DOI: 10.21437/Interspeech.2022-209. URL: https://www.isca-archive.org/interspeech_2022/zhang22_interspeech.html (visited on 2025-04-04).

Zhang, Junbo, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang (2021). *Speechocean762: An Open-Source Non-native English Speech Corpus For Pronunciation Assessment*. June 2, 2021. DOI: 10.48550/arXiv.2104.01378. arXiv: 2104.01378 [cs]. URL: http://arxiv.org/abs/2104.01378 (visited on 2025-04-04). Pre-published.

Zhao, Guanlong, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna (2018). "L2-ARCTIC: A Non-native English Speech Corpus". In: *Interspeech 2018*. Interspeech 2018. ISCA, Sept. 2, 2018, pp. 2783–2787. DOI: 10.21437/Interspeech.2018-1110. URL: https://www.isca-archive.org/interspeech_2018/zhao18b_interspeech.html (visited on 2025-04-04).