UPPSALA
UNIVERSITET

My thesis

Peter Cady

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

August 31, 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

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 automatic speech recognition (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 Mispronunciation Detection and Diagnosis (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 supposedly language independent word-based n-gram approaches to languages rests on its 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. How a language is communicated through script is also a point of complication for NLP research, with Manohar et al. (2024) implicating standard normalization praxis when comparing ASR models as artificially inflating scores of languages using Indic scripts. Correctly segmenting

continuous-script languages like Chinese is another area of ongoing research with clear implications for downstream performance. cut this. text work is not directly relevant to my project, and it doesn't illuminate any particular point.. By explicitly mobilizing linguistic knowledge already painstakingly gathered by linguistic typologists, we can identify where languages agree, where they differ, and hopefully identify implicit, ungeneralizable assumptions underpinning our approaches earlier in the design stage.

[11] language disparity in lang tech. Despite the claims of language-agnostic systems often touted by proponents of emerging Artificial Intelligence (AI) technologies, these systems have often 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 have no footprint in emerging language technologies (Joshi et al., 2021). In the past, building neural network (NN)-based language technologies has 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 such issues(see Magueresse 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 on page 10, which outlines the states and challenges faced by languages in resource terms in the digital space, and how dominant a small group of languages are within it. 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 outlined 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 Gaeltachtaí*. Each of the three main areas (i.e. Ulster, Connacht, and Munster 2.1) 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 same

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 cultural community online. Lacking in 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).

access to resources. It should be noted that the pretraining/finetuning paradigm of recent massive multilingual models does mitigate this demand for data somewhat 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.which survey references do I want to use?

[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., n.d.). Despite the state of Irish as a first language (L1), it is comparatively strong as a L2 broinNewUrbanIrish. 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., n.d., 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, n.d.; Gabriele, n.d.; Snesareva, 2016; Stenson, 2020). Since the Roman alphabet

Figure 2.1: Area of the Gaeltachtaí (Irish-speaking areas of Ireland) colored in green

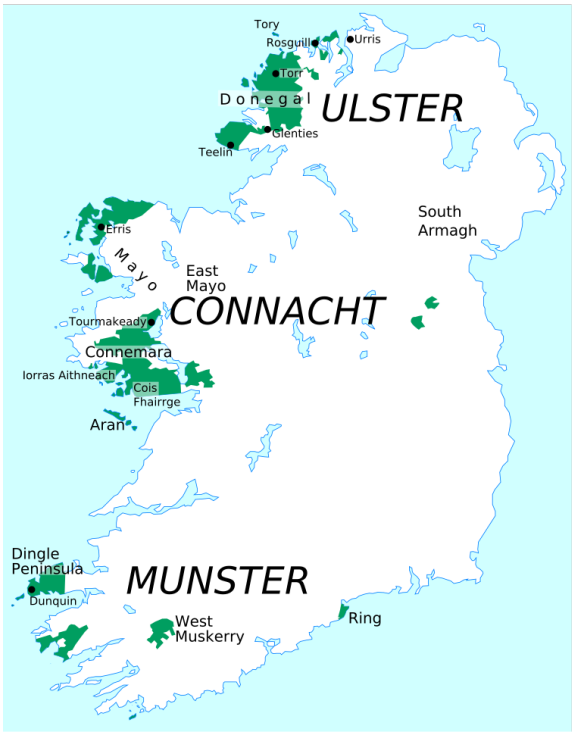

Figure 2.2: The original uploader was Angr at English Wikipedia. - Transferred from en.wikipedia to Commons., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=3532749

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') flanking a consonant denote *palatalisation*, while *broad* vowels ('a', 'o', and 'u') mark *velarisation* (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* (/ˈʊrʸuː/), 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 byBroin (n.d.) reveals that realisations of phonological like those above are 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 more attainable between gaeltacht and urban speakers.

Gabriele (n.d.) english influence on palatalization and velarization Snesareva (2016) how does english influence Irish spoken by dublin bilinguals? Broin (n.d.) differences

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

	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'

of irish between cities and gaeltacht Ní Chasaide et al. (n.d.) 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 at the beginning of this chapter. 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 (Hedderich et al., 2021). These data disparities can be categorized along several dimensions, such as those proposed by Hedderich et al. (2021) for NLP 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; Deng and Platt, 2014; Fiscus, 1997; Gitman et al., 2023, inter alia). Various methods of data augmentation 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 TTS-generated synthetic audio is used 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 (Rogerson-Revell, 2021), echoing calls for greater collaboration between pedagogical and technical experts when designing these systems.

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 on. This kind of individualized, explained 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 Automatic Speech Recognition

[16] general overview of ASR goal: transcribe speech to text not done, get good descriptors from **jurafskySpeechLanguageProcessing2025empty citation** and **yuAutomaticSpeechRecognition2**

[17] old asr models, hmm gnn 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).

[18] Description of Wav2vec2 model Conneau et al. (2020) Wav2vec2 XLSR Baevski, Schneider, et al. (2020) (vector quantized) wav2vec2 with learned discrete representations One popular ASR models used in extracting phonemes from raw acoustic data is currently *wav2vec 2.0*, a self-supervised framework which is conceptually simpler than other leading models. After pre-training on unlabeled data, it can be fine-tuned on a much smaller set of labeled data, achieving State-of-the-Art (SotA) 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.

[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

2.5 Pronunciation & Mispronunciation Detection and Diagnosis (MDD)

To realize the often untapped potential of CAPT and lay the groundwork for actionable feedback, the ASR 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. We must first clarify what we mean by pronunciation, starting with an abstraction of the human speech apparatus as a collection of subsystems which can emit signals over parallel channels (Engstrand, 2004). Although speech

is a continuous signal, we can map symbols of speech sounds–phones–onto its subsegments. 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 I use here, or consonantal as in Semitic scripts(**jurafskySpeechLanguageProcessing2025**). For our purposes, we will narrow our investigation of MD to 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.

[20] difference in objectives To

[21] classical gop approaches Research into MDD gained significant momentum in the 1990's to early 2000's, centering on hidden Markov model (HMM)-Gaussian Mixture Model (GMM) based ASR systems. Early approaches pronunciation scoring using these systems typically relied on log-likelihood scores and log-posterior scores (S. M. Witt, n.d.) the latter of which eventually eclipsing the former due to its higher correlation with human assessments expand this section. explain the log-likelihood approach and posterior with equations. it will be relevant in related works when tying it to wav2vec2 outputs. 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). S. M. Witt (2000) GOP dissertation Sudhakara et al. (2019) improved GOP Fu et al. (2021) data augmentation for MDD Kheir (n.d.) MDD with hmm gnn for arabic Kheir, Ali, et al. (2023) review of MDD capt systems Kheir, Chowdhury, et al. (2023) Multilingual MDD framework

[22] E2E ASR approaches More recently, end-to-end (E2E) NN-based based systems like Wav2Vec2 have gained traction as promising alternatives to more traditional HMM-GMM approaches. Logit outputs of these models are analogous to the posterior probabilities derived from HMMs, providing probabilities of a phoneme given the speech segment, but suffers from issues of 'peakiness' and overconfidence which complicate their interpretability as direct analogues to hmm-based posterior probabilities (**parikhEvaluatingLogitBasedGOP2025**) touches on confidence and peakiness(Zeyer et al., 2021) peakiness

Alrashoudi et al. (2025) MDD for arabic with transformers Kim et al. (2022) focus more on pron scoring instead of MDD 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

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 ASR for MDD

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

[24] MDD scoring

3.2 Ensembling

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

[25] describe the general strategy of model ensembling Fiscus (1997) ROVER ensembling 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 TTS data augmentation

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

[26] 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)

khareLowResourceASR2021empty citation 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)

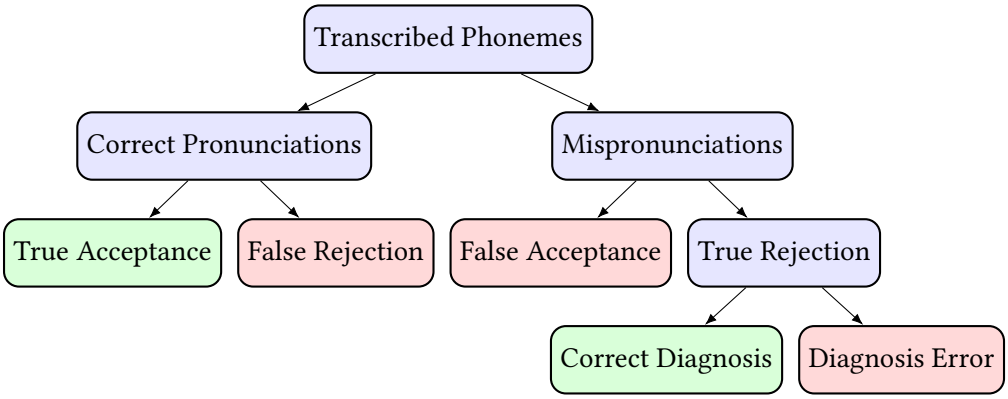

Figure 3.1: Evaluation hierarchy of phoneme level MD

4 Methods & Materials

In the following section, we will describe 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 dataset bootstrapping with a TTS system to create the data required to train a model directly for MD. Leading this will be an overview of the model framework common to both of these experiments: Wav2vec 2.0.

4.1 General Experiment Setup

Connect the low resource dimensions and approaches to my contributions.

[27] teanglann, timit, tts

[28] preprocessing, cleaning, etc.

[29] base model description

4.2 Experiment 1: Monolingual Ensembling

[30] outline first experiment with model ensemble. Our first experiment starts with the intuition described above using two expert systems: one Irish and one English. Confidence values associated with each models phonemic outputs are compared to each other to reach a combined phonetic output. 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

[31] 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]. To evaluate effectiveness of the system for MD, a manually annotated datset was prepared from a small subsection of the Mozilla Common Voice dataset Krishenbaum (n.d.) use is ascii is used (it's not) Měchura (n.d.) introduction to grammar part of teanglann (maybe not relevant)

[32] 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

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

of dialects. The full dataset consists of 6300 sentences from 630 speakers, so ten sentences per speaker. This Ardila et al. (2020) common voice corpus for L2 speech

[33] teanglann The Dictionary and Language Library is an online resource developed by Foras na Gaeilge[2] in conjunction with the New English-Irish Dictionary. Among the resources available are The Pronunciation Database, which contains recordings of individual words spoken by native speakers from three major dialects: Connacht, Ulster, and Munster. The recordings for Ulster 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. am i using MFA or the dict? The dictionary of words to scrape was derived from an Ulster Irish Grapheme-to-Phoneme (G2P) file generously provided by Jim O' Regan. This was used for canonical Ulster phonetic annotation, and could be combined to form a dataset of words, recordings, and phonetic annotations in IPA for a combined dataset 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

4.2.2 Models

[34] which wav2vec2 variant I used

4.3 Experiment 2: Synthetic Mispronunciation Data

4.3.1 Data

zhang2022l2genempty citation synthetic data generation

4.3.2 Models

4.4 Evaluation

[35] overview, skyline, baseline **graves2006connectionistempty citation** ctc loss

[36] data used

[37] 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 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 an Montreal Forced Aligner (MFA) G2P model trained on Ulster Irish to speed up the process ask about where the g2p came from, what to reference

[38] Common voice For evaluating the quality of the MD system, 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[3]. As in their work, this dataset will

[2] 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/

[3] given its nature as a crowd-sourced dataset, it is not certain this is still the case for the Irish portion of Common Voice

be used as the basis for testing the current experiment, as it fits the purpose of evaluating our system's effectiveness on L2 speech. 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. 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)

[39] 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

[40] protocol, why does this enable comparison? (bridge between methods, materials, and results)

5 Results

[41] detail the results as it pertains to the evaluation framework: CER

6 Discussion

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

[43] 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

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

[45] continue furhat explanation

[46] 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

[47] 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

[48] 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

[49] 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

[50] General Ethical considerations for current work and possible extensions

[51] Ethical considerations for Minority languages

9 AI Tools

[52] 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).

Alrashoudi, Norah, Hend Al-Khalifa, and Yousef Alotaibi (2025). "Improving Mispronunciation Detection and Diagnosis for Non- Native Learners of the Arabic Language". *Discover Computing* 28.1 (Jan. 6, 2025), p. 1. ISSN: 2948-2992. DOI: 10.1007/s10791-024-09489-8. URL: https://link.springer.com/10.1007/s10791-024-09489-8 (visited on 2025-06-16).

Amrate, Moustafa and Pi-hua Tsai (2025). "Computer-Assisted Pronunciation Training: A Systematic Review". *ReCALL* 37.1 (Jan. 2025), pp. 22–42. ISSN: 0958-3440, 1474-0109. DOI: 10.1017/S0958344024000181. URL: https://www.cambridge.org/core/product/identifier/S0958344024000181/type/journal_article (visited on 2025-04-04).

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.

Arnbjörnsdóttir, Birna, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton, eds. (2022). *Intelligent CALL, Granular Systems and Learner Data: Short Papers from EUROCALL 2022*. 1st ed. Research-publishing.net, Dec. 12, 2022. ISBN: 978-2-38372-015-7. DOI: 10.14705/rpnet.2022.61.9782383720157. URL: https://research-publishing.net/book?10.14705/rpnet.2022.61.9782383720157 (visited on 2025-04-04).

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.

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, Henry Zhou, Abdelrahman Mohamed, and Michael Auli (2020). *Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. Oct. 22, 2020. DOI: 10.48550/arXiv.2006.11477. arXiv: 2006.11477 [cs]. URL: http://arxiv.org/abs/2006.11477 (visited on 2025-04-04). Pre-published.

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.

Ballier, Nicolas, Adrien Méli, Maelle Amand, and Jean-Baptiste Yunès (n.d.). "Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech: A Case Study with French Learners of English" ().

Bannò, Stefano and Marco Matassoni (2022). *Proficiency Assessment of L2 Spoken English Using Wav2vec 2.0*. Oct. 24, 2022. DOI: 10.48550/arXiv.2210.13168. arXiv: 2210.13168 [cs]. URL: http://arxiv.org/abs/2210.13168 (visited on 2025-04-04). Pre-published.

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).

Boulianne, Gilles (2022). "Phoneme Transcription of Endangered Languages: An Evaluation of Recent ASR Architectures in the Single Speaker Scenario". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, pp. 2301–2308. DOI: 10.18653/v1/2022.findings-acl.180. URL: https://aclanthology.org/2022.findings-acl.180 (visited on 2025-04-04).

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

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.

Chen, Lei, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian (2018). "End-to-End Neural Network Based Automated Speech Scoring". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB: IEEE, Apr. 2018, pp. 6234–6238. ISBN: 978-1-5386-4658-8. DOI: 10.1109/ICASSP.2018.8462562. URL: https://ieeexplore.ieee.org/document/8462562/ (visited on 2025-04-04).

Chen, Nancy F. and Haizhou Li (2016). "Computer-Assisted Pronunciation Training: From Pronunciation Scoring towards Spoken Language Learning". In: *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Jeju, South Korea: IEEE, Dec. 2016, pp. 1–7. ISBN: 978-988-14768-2-1. DOI: 10.1109/APSIPA.2016.7820782. URL: http://ieeexplore.ieee.org/document/7820782/ (visited on 2025-04-04).

Collins, Naoise (2021). “Situated Immersive Gaming Environments for Irish Language Learning”. DOI: 10.21427/JKMJ-XM34. URL: https://arrow.tudublin.ie/tourdoc/35 (visited on 2025-04-04).

Collins, Naoise, Dr Brian Vaughan, Dr Charlie Cullen, and Dr Keith Gardner (2019). “GaeltechVR: Measuring the Impact of an Immersive Virtual Environment to Promote Situated Identity in Irish Language Learning”. 12.3.

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*.

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).

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.

Eskenazi, Maxine (2009). “An Overview of Spoken Language Technology for Education”. *Speech Communication* 51.10 (Oct. 2009), pp. 832–844. ISSN: 01676393. DOI: 10.1016/j.specom.2009.04.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639309000673 (visited on 2025-04-04).

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.

Fu, Kaiqi, Jones Lin, Dengfeng Ke, Yanlu Xie, Jinsong Zhang, and Binghuai Lin (2021). *A Full Text-Dependent End to End Mispronunciation Detection and Diagnosis with Easy Data Augmentation Techniques*. Apr. 17, 2021. DOI: 10.48550/arXiv.2104.08428. arXiv: 2104.08428 [cs]. URL: http://arxiv.org/abs/2104.08428 (visited on 2025-04-04). Pre-published.

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 Edwards, Jette G. and Mary L. Zampini, eds. (2008). *Phonology and Second Language Acquisition*. Studies in Bilingualism. Amsterdam: John Benjamins Publishing Company. 1 p. ISBN: 978-90-272-4147-4 978-90-272-9139-4. DOI: 10.1075/sibil.36.

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.

Holmberg, Jörgen (n.d.). "Designing for Added Pedagogical Value" ().

Homa, Donald and Joan Cultice (n.d.). "Role of Feedback, Category Size, and Stimulus Distortion on the Acquisition and Utilization of Ill-Defined Categories" ().

Hosseini-Kivanani, Nina, Roberto Gretter, Marco Matassoni, and Giuseppe Daniele Falavigna (2021). *Experiments of ASR-based Mispronunciation Detection for Children and Adult English Learners*. Apr. 13, 2021. DOI: 10.48550/arXiv.2104.05980. arXiv: 2104.05980 [cs]. URL: http://arxiv.org/abs/2104.05980 (visited on 2025-04-04). Pre-published.

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed (2021). *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. June 14, 2021. DOI: 10.48550/arXiv.2106.07447. arXiv: 2106.07447 [cs]. URL: http://arxiv.org/abs/2106.07447 (visited on 2025-04-04). Pre-published.

Islam, Elaf, Chanho Park, and Thomas Hain (2023). "Exploring Speech Representations for Proficiency Assessment in Language Learning". In: *9th Workshop on Speech and Language Technology in Education (SLaTE)*. 9th Workshop on Speech and Language Technology in Education (SLaTE). ISCA, Aug. 18, 2023, pp. 151–155. DOI: 10.21437/SLaTE.2023-29. URL: https://www.isca-archive.org/slate_2023/islam23_slate.html (visited on 2025-04-04).

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.

Kheir, Yassine EL (n.d.). "Mispronunciation Detection with SpeechBlender Data Augmentation Pipeline." ().

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.

Kheir, Yassine El, Shammur Absar Chowdhury, and Ahmed Ali (2023). *L1-Aware Multilingual Mispronunciation Detection Framework*. Sept. 21, 2023. DOI: 10.48550/arXiv.2309.07719. arXiv: 2309.07719 [cs]. URL: http://arxiv.org/abs/2309.07719 (visited on 2025-04-04). Pre-published.

Kim, Eesung, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim (2022). *Automatic Pronunciation Assessment Using Self-Supervised Speech Representation Learning*. Apr. 8, 2022. DOI: 10.48550/arXiv.2204.03863. arXiv: 2204.03863 [eess]. URL: http://arxiv.org/abs/2204.03863 (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).

Kyriakopoulos, Konstantinos, Kate Knill, and Mark Gales (2018). "A Deep Learning Approach to Assessing Non-native Pronunciation of English Using Phone Distances". In: *Interspeech 2018*. Interspeech 2018. ISCA, Sept. 2, 2018, pp. 1626–1630. DOI: 10.21437/Interspeech.2018-1087. URL: https://www.isca-archive.org/interspeech_2018/kyriakopoulos18_interspeech.html (visited on 2025-04-04).

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" ().

Levis, John (2007). "COMPUTER TECHNOLOGY IN TEACHING AND RESEARCHING PRONUNCIATION". *Annual Review of Applied Linguistics* 27 (Mar. 2007). ISSN: 0267-1905, 1471-6356. DOI: 10.1017/S0267190508070098. URL: http://www.journals.cambridge.org/abstract_S0267190508070098 (visited on 2025-04-04).

Levis, John and Lucy Pickering (2004). "Teaching Intonation in Discourse Using Speech Visualization Technology". *System* 32.4 (Dec. 2004), pp. 505–524. ISSN: 0346251X. DOI: 10.1016/j.system.2004.09.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0346251X04000752 (visited on 2025-04-04).

Li, Kun, Xiaojun Qian, and Helen Meng (2017). "Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks". *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (Jan. 2017), pp. 193–207. ISSN: 2329-9290, 2329-9304. DOI: 10.1109/TASLP.2016.2621675. URL: http://ieeexplore.ieee.org/document/7752846/ (visited on 2025-04-04).

Li, Sheng, Lan Wang, and En Qi (2011). "The Phoneme-Level Articulator Dynamics for Pronunciation Animation". In: *2011 International Conference on Asian Language Processing*. 2011 International Conference on Asian Language Processing (IALP). Penang, Malaysia: IEEE, Nov. 2011, pp. 283–286. ISBN: 978-1-4577-1733-8. DOI: 10.1109/IALP.2011.13. URL: http://ieeexplore.ieee.org/document/6121521/ (visited on 2025-04-04).

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).

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.

Lyster, Roy, Kazuya Saito, and Masatoshi Sato (2013). "Oral Corrective Feedback in Second Language Classrooms". *Language Teaching* 46.1 (Jan. 2013), pp. 1–40. ISSN: 0261-4448, 1475-3049. DOI: 10.1017/S0261444812000365. URL: https://www.cambridge.org/core/product/identifier/S0261444812000365/type/journal_article (visited on 2025-04-04).

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.

Manohar, Kavya, Leena G. Pillai, and Elizabeth Sherly (2024). *What Is Lost in Normalization? Exploring Pitfalls in Multilingual ASR Model Evaluations*. Nov. 9, 2024. DOI: 10.48550/arXiv.2409.02449. arXiv: 2409.02449 [cs]. URL: http://arxiv.org/abs/2409.02449 (visited on 2025-04-04). Pre-published.

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

Mihalicek, Vedrana and Christin Wilson, eds. (2012). *Language Files: Materials for an Introduction to Language and Linguistics*. 11th edition. Taipei: Bookman Books. ISBN: 978-957-445-461-7.

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" ().

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 (n.d.). "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 Commu-

nity". 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).

Ní Chiaráin, Neasa (2022). "An Corpas Cliste: Creating a Learner Corpus for Irish from a New, Purpose-Built iCALL Platform". In: *Intelligent CALL, Granular Systems and Learner Data: Short Papers from EUROCALL 2022*. Ed. by Birna Arnbjörnsdóttir, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton. 1st ed. Research-publishing.net, Dec. 12, 2022, pp. 297–301. ISBN: 978-2-38372-015-7. DOI: 10.14705/rpnet.2022.61.1474. URL: https://research-publishing.net/manuscript?10.14705/rpnet.2022.61.1474 (visited on 2025-04-04).

Ní Chiaráin, Neasa and Ailbhe Ní Chasaide (2016). "The Digichaint Interactive Game as a Virtual Learning Environment for Irish". In: *CALL Communities and Culture – Short Papers from EUROCALL 2016*. Ed. by Salomi Papadima-Sophocleous, Linda Bradley, and Sylvie Thouësny. Research-publishing.net, Dec. 18, 2016, pp. 330–336. ISBN: 978-1-908416-44-5. DOI: 10.14705/rpnet.2016.eurocall2016.584. URL: https://research-publishing.net/manuscript?10.14705/rpnet.2016.eurocall2016.584 (visited on 2025-04-04).

Ní Chiaráin, Neasa and Ailbhe Ní Chasaide (2018). "An Scéalaí: Synthetic Voices for Autonomous Learning". In: Taalas, Peppi, Juha Jalkanen, Linda Bradley, and Sylvie Thouësny. *Future-Proof CALL: Language Learning as Exploration and Encounters – Short Papers from EUROCALL 2018*. Research-publishing.net, Dec. 8, 2018, pp. 230–235. ISBN: 978-2-490057-22-1. DOI: 10.14705/rpnet.2018.26.842. URL: https://research-publishing.net/manuscript?10.14705/rpnet.2018.26.842 (visited on 2025-04-04).

Ní Chiaráin, Neasa, Oisín Nolan, Madeleine Comtois, Neimhin Robinson Gunning, Harald Berthelsen, and Ailbhe Ni Chasaide (2022). "Using Speech and NLP Resources to Build an iCALL Platform for a Minority Language, the Story of An Scéalaí, the Irish Experience to Date". In: *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages. Dublin, Ireland: Association for Computational Linguistics, pp. 109–118. DOI: 10.18653/v1/2022.computel-1.14. URL: https://aclanthology.org/2022.computel-1.14 (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).

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.

Pisoni, David B and Robert E Remez (n.d.). "The Handbook of Speech Perception" ().

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).

Saito, Kazuya and Luke Plonsky (2019). "Effects of Second Language Pronunciation Teaching Revisited: A Proposed Measurement Framework and Meta-Analysis". *Language Learning* 69.3 (Sept. 2019), pp. 652–708. ISSN: 0023-8333, 1467-9922. DOI: 10.1111/lang.12345. URL: https://onlinelibrary.wiley.com/doi/10.1111/lang.12345 (visited on 2025-04-04).

Schmidt, Richard (2012). "Chapter 2. Attention, Awareness, and Individual Differences in Language Learning". In: *Perspectives on Individual Characteristics and Foreign Language Education*. Ed. by Wai Meng Chan, Kwee Nyet Chin, Sunil Bhatt, and Izumi Walker. DE GRUYTER, Sept. 13, 2012, pp. 27–50. ISBN: 978-1-61451-095-6 978-1-61451-093-2. DOI: 10.1515/9781614510932.27. URL: https://www.degruyter.com/document/doi/10.1515/9781614510932.27/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).

Sheen, YoungHee (2004). "Corrective Feedback and Learner Uptake in Communicative Classrooms across Instructional Settings". *Language Teaching Research* 8.3 (July 2004), pp. 263–300. ISSN: 1362-1688, 1477-0954. DOI: 10.1191/1362168804lr146oa. URL: https://journals.sagepub.com/doi/10.1191/1362168804lr146oa (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).

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.

Strik, Helmer, Khiet Truong, Febe De Wet, and Catia Cucchiarini (2009). "Comparing Different Approaches for Automatic Pronunciation Error Detection". *Speech Communication* 51.10 (Oct. 2009), pp. 845–852. ISSN: 01676393. DOI: 10.1016/j.specom.2009.05.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639309000715 (visited on 2025-04-04).

Sudhakara, Sweekar, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh (2019). "An Improved Goodness of Pronunciation (GoP) Measure for Pronunciation Evaluation with DNN-HMM System Considering HMM Transition Probabilities". In: *Interspeech 2019*. Interspeech 2019. ISCA, Sept. 15, 2019, pp. 954–958. DOI: 10.21437/Interspeech.2019-2363. URL: https://www.isca-archive.org/interspeech_2019/sudhakara19_interspeech.html (visited on 2025-04-04).

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).

Thomson, Ron and Tracey Derwing (2014). "The Effectiveness of L2 Pronunciation Instruction: A Narrative Review". *Applied Linguistics* 2014 (Dec. 8, 2014), pp. 1–20. DOI: 10.1093/applin/amu076.

VanPatten, Bill (n.d.). *Theories in Second Language Acquisition.*

Volodina, Elena, Gintarė Grigonytė, Ildikó Pilán, Kristina Nilsson Björkenstam, and Lars Borin (2016). "Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition". In: *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition.*

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, S.M and S.J Young (2000). "Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning". *Speech Communication* 30.2–3 (Feb. 2000), pp. 95–108. ISSN: 01676393. DOI: 10.1016/S0167-6393(99)00044-8. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639399000448 (visited on 2025-04-04).

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.

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).

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).