Automatic Assessment of L2 Speech Intelligibility and Pronunciation

Xing Wei

Centre for Language Studies

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Automatic Assessment of L2 Speech Intelligibility and Pronunciation

Xing Wei

Radboud Universiteit

The research reported in this thesis has been carried out under the auspices of the Centre for Language Studies of Radboud University and the China Scholarship Council.

Automatic Assessment of L2 Speech Intelligibility and Pronunciation
Xing Wei

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Proefschrift AIO | Annelies Lips
Cover: Proefschrift AIO | Guntra Laivacuma
Printing: DPN Rikken/Pumbo

ISBN: 9789465152967
DOI: 10.54195/9789465152967
Free download at: https://doi.org/10.54195/9789465152967

RADBOUD UNIVERSITY PRESS

Automatic Assessment of L2 Speech Intelligibility and Pronunciation

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 25 augustus 2026
om 14.30 uur precies

door

Xing Wei
geboren op 19 juni 1989
te Wuhan (China)

Promotor:
Dr. W.A.J. Strik

Copromotoren:
Dr. C. Cucchiarini
Prof. dr. R.W.N.M. van Hout

Manuscriptcommissie:
Prof. dr. M.T.C. Ernestus
Prof. dr. J. Colpaert (Universiteit Antwerpen, België)
Prof. dr. ir. H. Van hamme (KU Leuven, België)
Prof. dr. H. van de Velde (Universiteit Utrecht)
Dr. B. Schuppler (Technische Universität Graz, Oostenrijk)

Automatic Assessment of L2 Speech Intelligibility and Pronunciation

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Tuesday, August 25, 2026
at 2.30 pm

by

Xing Wei
born on June 19, 1989
in Wuhan (China)

Supervisor:
Dr. W.A.J. Strik

Co-supervisors:
Dr. C. Cucchiarini
Prof. dr. R.W.N.M. van Hout

Manuscript Committee:
Prof. dr. M.T.C. Ernestus
Prof. dr. J. Colpaert (University of Antwerp, Belgium)
Prof. dr. ir. H. Van hamme (KU Leuven, Belgium)
Prof. dr. H. van de Velde (Utrecht University)
Dr. B. Schuppler (Graz University of Technology, Austria)

Table of contents


Hello
你好

Chapter 1

Introduction

In an era defined by increasing globalization and mobility, the ability to communicate across linguistic and cultural boundaries has shifted from a specialized skill to an essential component of social and professional participation. The global spread and functional importance of major world languages, such as English, have contributed to sustained growth in second language (L2) learning worldwide (Cook, 2016; Crystal, 2003; Warschauer & Healey, 1998). Among the various competencies required for successful L2 acquisition, spoken proficiency remains particularly central and simultaneously one of the most challenging to develop (Bygate, 1987; Levelt, 1993; Luoma, 2004). Speaking demands not only linguistic knowledge, but also the real-time processing and interactional skills that underpin communicative competence (Canale & Swain, 1980). Effective oral communication is the cornerstone of meaningful interaction, enabling learners to participate in social, academic, and professional conversations (Celce-Murcia et.al., 2010; Jack & Richards, 2008).

A persistent difficulty in conventional classroom instruction is the limited opportunity for learners to receive individualized and timely feedback on their spoken performance. Large class sizes, inadequate opportunities for interaction, and restricted instructional time often prevent teachers from providing the consistent and targeted responses required to support improvements in pronunciation and other components of oral proficiency (Derwing & Munro, 2005; Ellis, 2009; Lyster & Saito, 2010). In response to these pedagogical constraints, research within Computer-Assisted Language Learning (CALL) has increasingly investigated technological solutions designed to supplement or extend teacher instruction, particularly in contexts where individualized guidance is difficult to deliver (Chapelle, 2001; Hubbard, 2008; Levy & Stockwell, 2013). Automated CALL systems for spoken language assessment can offer forms of feedback that are rarely achievable in traditional classrooms, providing timely, objective, and scalable evaluations that facilitate independent practice and sustained learning (Eskenazi, 2009; Neri et al., 2008; Witt, 2012). This synthesis of pedagogy and technology reflects the multidisciplinary essence of CALL, which seeks to integrate diverse theoretical perspectives to optimize the language learning environment (Levy, 1997). This perspective is also consistent with recent work on Smart CALL, which emphasizes personalization, contextualization, and socialization as important dimensions of technology-supported language learning (Colpaert & Stockwell, 2022). In the context of automatic speech assessment, personalization is particularly relevant, as learners may benefit from feedback that is adapted to their own pronunciation patterns and intelligibility-related difficulties.

Within this research area, two domains have received particular attention: pronunciation assessment and intelligibility assessment. Pronunciation assessment concerns the systematic evaluation of how accurately learners produce the sound patterns of the target language. Since pronunciation is a complex productive skill, its assessment typically considers both segmental features, such as vowels and consonants, and suprasegmental features, including stress, rhythm, and intonation, all of which contribute to communication effectiveness (Hahn, 2004; Jenkins, 2000; Levis, 2005). Intelligibility assessment, by contrast, focuses on how well listeners can understand a learner's spoken output. As a listener-based construct central to communication, intelligibility reflects the extent to which meaning is successfully conveyed. While distinct from comprehensibility, which measures the ease of understanding, intelligibility has long been regarded as a key outcome in spoken language learning (Munro & Derwing, 1995; Thomson & Derwing, 2015).

Despite the promise of novel technologies for automated CALL systems, the development of such systems remains challenging, primarily due to the inherent characteristics of L2 speech. Non-native speech typically exhibits substantially greater variability than native speech, both within and across speakers (Hecht and Mulford, 1982; Major, 2001). Much of this variability stems from systematic L1 influence, which causes predictable transfer patterns at the segmental and prosodic levels (Best et al., 2007; Flege, 1987; Gass et al., 2020; Lado, 1957). Additional sources of variation include learners' proficiency levels (Tremblay, 2001) and individual differences in learning experience and strategy use. These factors jointly produce a wide range of acoustic-phonetic realizations that differ from native norms. As a result, automatic systems, most of which are trained primarily on large native speech datasets, often have trouble adapting to L2 productions, leading to degraded recognition and reduced reliability in assessment (Benzeghiba et al., 2007; Witt, 2012).

To address these critical issues, this thesis aims to advance the automatic assessment of L2 speech intelligibility and pronunciation. It investigates novel methods to enhance the robustness and diagnostic capacity of assessment models through three primary avenues of research. First, it explores the use of diverse and rich feature representations, moving beyond traditional acoustic features to incorporate linguistically grounded information. Second, it delves into the fundamental nature of intelligibility itself, comparing different methods of measurement and developing models that can predict human perceptual judgments from the speech signal. Third, it examines and applies advanced modeling architectures, including state-of-the-art (SOTA) end-to-end (E2E) based models, to better capture the complex patterns of L2 speech.

1.1 Speech intelligibility assessment

While pronunciation assessment concentrates on the accuracy of speech at a detailed, phonetic level, speech intelligibility addresses a more fundamental and communication-focused question: How well can a listener understand what is being said? Intelligibility serves as a comprehensive measure of effective communication and is often viewed as a primary goal in L2 instruction, as it directly relates to the learner's ability to convey meaning successfully in real-world interactions (Derwing & Munro, 2005; Isaacs & Trofimovich, 2012). Automatically evaluating intelligibility is therefore an essential part of a comprehensive CALL system, offering a measure that complements the more prescriptive, form-focused assessment obtained from pronunciation analysis.

Assessing speech intelligibility, however, is a complex task. It is not an inherent property of the speech signal itself; rather, it arises from the dynamic interaction between the speaker, the listener, and the context (Hazan & Markham, 2004). Research in this domain has traditionally been divided into two main trends: subjective assessment, which depends on the perceptual judgments of human listeners, and objective assessment, which aims to predict these judgments using computational models. The difference between these approaches, along with their respective advantages and challenges, plays a critical role in advancing methods for measuring intelligibility.

1.1.1 Subjective assessment methods

Subjective methods, which are based on judgments by human raters to evaluate speech, are regarded as the "gold standard" for measuring intelligibility because they directly capture the perceptual outcome of communication (Munro & Derwing, 1995). The primary approaches for subjective assessment fall into two categories. The first and most straightforward method is orthographic transcription, where listeners are asked to write down what they hear a speaker say (Bent & Bradlow, 2003). The intelligibility score is then determined as the percentage of words or phonemes correctly transcribed, providing a clear measure of information transfer. The second type of subjective assessment uses scalar rating scales, often employed to measure the related construct of perceived ease of understanding (Derwing & Munro, 1997). Listeners rate the speech on a numerical continuum. Common methods include the application of the Likert scale, which uses a fixed set of discrete points (e.g., 1 to 7) (Likert, 1932), and the Visual Analogue Scale (VAS), where listeners mark a position on a continuous horizontal line anchored by descriptive endpoints (Wewers & Lowe, 1990).

While these methods are foundational, they face notable practical challenges. Orthographic transcription is very time-consuming and labor-intensive for both listeners and researchers, making it impractical for large-scale assessments or real-time feedback (Gass & Varonis, 1984). Scalar ratings, although faster, introduce inherent subjectivity. Ratings can be affected by listener-related factors, such as their tolerance for accented speech or their L1 background (Kang et al., 2018). To achieve consistent measures with either method, it is common to average scores from a panel of multiple listeners, which further increases cost and complexity (Shrout & Fleiss, 1979). This high cost and limited scalability pose a significant challenge, driving the pursuit of automated, objective alternatives.

1.1.2 Objective assessment methods

Given the significant practical limitations of subjective assessment methods, a key goal in CALL research is to develop objective measures that can automatically predict human intelligibility judgments from the acoustic signal (Ma et al., 2009; Xue et al., 2019; Karbasi & Kolossa, 2022). An early intuitive approach was to use an ASR system as a substitute for a human listener. The Word Error Rate (WER) generated by the ASR system could theoretically serve as an inverse indicator of intelligibility (Fontan et al., 2017; Gallardo et al., 2017; Martinez et al., 2022). A lower WER would theoretically correspond to higher intelligibility. However, this approach has proven problematic. As will be discussed later in section 1.2.1, standard ASR systems are often inaccurate when processing L2 speech. More importantly, the types of errors causing ASR system failure do not always match those that would impair a human listener. As a result, many studies have found a weak and inconsistent correlation between the ASR-WER system and human intelligibility scores, limiting the practical usefulness of this direct ASR-based approach (Van Doremalen et al., 2010; Gutz et al., 2022; Inceoglu et al., 2023; Kim et al., 2024).

A more promising and well-studied research direction focuses on developing statistical or machine learning models that predict human-rated intelligibility scores based on a set of acoustic-phonetic features extracted from speech signals (Xue et al., 2021). This method aims to identify and measure the specific acoustic cues that human listeners rely on when assessing intelligibility. Researchers have explored a wide range of features, including those related to temporal aspects like articulation rate and pause duration (Cucchiarini et al., 2002; Kormos & Dénes, 2004; Trofimovich & Baker, 2006), as well as vowel space clarity, such as the Vowel Space Area (Bradlow et al., 1996). The main methodological challenge is creating a predictive model from this large, high-dimensional, and often highly correlated set of features. To overcome this, researchers have used feature selection and dimensionality reduction

techniques, such as Principal Component Analysis (PCA) (Fewzee & Karray, 2012), or more advanced regression methods like LASSO (Tibshirani, 1996) and Elastic Net (Zou & Hastie, 2005). Although these approaches show great promise, a key challenge remains the lack of systematic comparisons of different feature selection and modeling strategies for L2 speech. There is a clear need for studies that rigorously identify the most predictive acoustic-phonetic features and utilize robust statistical methods to develop reliable objective models of intelligibility.

1.2 Automatic pronunciation assessment

Automatic pronunciation assessment, a core component of modern CALL systems, is dedicated to the evaluation of L2 learners' spoken productions. Its primary function is often called Pronunciation Error Detection (PED) or referred to as Mispronunciation Detection and Diagnosis (MDD) in some literature (Zhang et al., 2020). It mainly involves identifying deviations from a native-like pronunciation in sounds, words, or phrases, while offering detailed diagnoses of the errors. This diagnostic capability is crucial for providing learners with targeted, actionable feedback that can improve their practice and speed up their phonological progress (Cucchiarini et al., 2009). From a pedagogical and psychological perspective, the provision of such detailed diagnostic feedback is closely linked to learner motivation. As noted by Colpaert (2024), the rationale behind language learning is deeply rooted in how learners perceive progress and utility; specific, actionable feedback can enhance self-efficacy and sustain the psychological drive necessary for long-term pronunciation improvement. The entire field is based on the technological foundation of ASR over recent decades, and its development has been about adapting and expanding ASR techniques to meet the specific needs of L2 speech assessment (Kheir et al., 2023).

Research in this field is mainly shaped by progress in three core, interconnected areas. These include the data used to train models, the features chosen to represent the speech signal, and the modeling strategies or architectures implemented. The areas of data, feature representations, and architectures are discussed below.

1.2.1 The challenge of data scarcity and variability

The performance of any data-driven speech technology is fundamentally constrained by the quality and quantity of the data it is trained on. For L2 pronunciation assessment, this creates a twofold challenge. The first is the persistent lack of appropriately annotated L2 corpora. The second is the significant variability inherent in non-native speech.

State-of-the-art ASR systems for native speech are routinely trained on thousands of hours of transcribed audio from publicly available corpora like Librispeech (Panayotov et al., 2015) or TIMIT (Garofolo et al., 1993). Such extensive data resources are a luxury rarely available for L2 learners. The process of collecting, transcribing, and, most importantly, phonetically annotating L2 speech with detailed error labels is a labor-intensive, time-consuming, and costly effort that requires significant phonetic expertise (Cucchiarini & Strik, 2017). This data bottleneck has been a well-documented and ongoing obstacle to developing robust and generalizable CALL systems (Eskenazi, 1999; Chen & Li, 2016).

The inherent variability in L2 speech further exacerbates this data issue. An assessment model trained on L2 speech from learners with a specific L1, such as the CU-CHLOE corpus of Cantonese English described by Meng et al. (2007), may perform poorly when evaluating speakers from a different L1 background, as their error patterns, driven by L1-L2 phonological contrasts, can be drastically different (Flege, 1995; Tu et al., 2018). Additionally, a learner's pronunciation is not fixed. It changes as their proficiency improves, with the frequent, salient errors of a beginner often being replaced by more subtle distortion errors in an advanced learner (Han, 2004). Efforts to develop dedicated L2 corpora, such as the L2-ARCTIC corpus (Zhao et al., 2018), are essential, but require significant resources and cannot cover all possible L1-L2 pairings.

In response to these long-standing challenges, researchers have developed and refined a range of data-driven methods. A common technique for artificially increasing the size and diversity of training data is data augmentation. Typical methods include adding various types of noise (Schuller et al., 2009), simulating reverberation (Ko et al., 2017), and applying vocal tract length perturbation (VTLP) to simulate speaker variability (Jaitly & Hinton, 2013). Among these, SpecAugment has proven particularly effective for ASR; by masking contiguous regions of frequency channels and time steps in the log-mel spectrogram, it encourages the model to learn more robust and invariant representations (Park et al., 2019). However, a key limitation of these approaches is their inability to simulate the nuanced, linguistically structured phonetic deviations seen in L2 speech. More advanced techniques aim to synthesize L2-accented speech (Rosenberg et al., 2019), but such methods remain challenging to control and validate.

Given the relative abundance of native speech corpora, a substantial amount of research has focused on adapting models trained exclusively on native speech. One of the most traditional and influential approaches is the Goodness of Pronunciation (GOP) method developed by Witt and Young in 2000. In this framework, an ASR

system trained on a native speech dataset is used to perform forced alignment, and the acoustic model then computes a goodness-of-fit score for each aligned phone segment. This score is typically defined as the log posterior probability of the canonical phone (Hu et al., 2015b), and it is often normalized by the posteriors of other competing phones to improve its discriminative power (Van Doremalen et al., 2013; Lo et al., 2020; Strik et al., 2009). Despite its widespread use, the effectiveness of GOP is fundamentally limited by the acoustic mismatch between the native training data and the non-native test data (Strik et al., 2007). This mismatch can result in a high rate of false rejections (Kanters et al., 2009; Zhao et al., 2012).

A more advanced strategy is transfer learning, where a model pre-trained on a large native speech dataset is fine-tuned with a smaller amount of L2 speech data. This enables the model to learn general acoustic-phonetic representations from the source data and adapt them to the specific characteristics of L2 speech. This method has proven highly effective in bridging the gap between native and non-native acoustic spaces (Huang et al., 2017; Nazir et al., 2019). Transfer learning principles can also be applied across different languages. Cross-lingual modeling leverages data from a high-resource language to improve systems for a low-resource language (Duan et al., 2019; Joshi et al., 2015). Multilingual models, trained jointly on data from multiple languages, learn to construct a shared phonetic space that remains effective across linguistic differences (Huang et al., 2013; Matassoni et al., 2018). This makes them especially suitable for L2 applications.

While the approaches discussed above clearly help address data scarcity, their application with pluricentric languages remains underexplored. These languages present additional challenges to ASR-based assessment because they include multiple standard varieties that differ systematically in phonetic and phonological features, with training resources being distributed unevenly among them (Clyne, 1992). Previous research has revealed notable differences across varieties in languages like English, Spanish, and Dutch (Kloots et al., 2006; van de Velde et al., 2010). As a result, learners working with a non-dominant variety, such as Flemish Dutch, often face more severe data limitations than those using the dominant standard. Studies like those of Yilmaz et al. (2016) suggest that combining data from different varieties of the Dutch language for modeling can aid in pathological speech assessment. However, it remains unclear whether cumulating resources from a dominant variety will improve or hinder ASR and MDD performance for less-resourced varieties in L2 contexts involving pluricentric languages. This uncertainty highlights an essential methodological issue related to selecting data and adapting models for pronunciation assessment in these languages.

1.2.2 The evolution of feature representations in pronunciation assessment

The features used to represent the speech waveform are crucial for both the performance and the diagnostic utility of an ASR-based pronunciation assessment system. These features serve as the input to the modeling architecture and therefore determine what information the system can extract from the signal. Research in this area has advanced from low-level acoustic descriptors to representations that incorporate more explicit linguistic structure designed to capture phonetic detail relevant for assessment.

Early approaches in ASR and ASR-based CALL systems depended on features derived from short-term spectral analysis. These representations aimed to capture perceptually meaningful information while providing a compact signal description. Mel-Frequency Cepstral Coefficients (MFCCs) are among the most widely used acoustic features in speech processing (Davis & Mermelstein, 1980). The original formulation produced a mel-frequency cepstrum from log-mel filterbank outputs via cepstral analysis. In later, now-standard implementations, a Discrete Cosine Transform (DCT) is applied to the log-mel energies to produce decorrelated, compact cepstral coefficients (e.g., Huang et al., 2001; Rabiner & Juang, 1993; Young et al., 2002). Other well-established representations include Perceptual Linear Prediction (PLP) coefficients (Hermansky, 1990) and log-mel filterbank energies (FBanks). Notably, FBanks, which are the log-mel filterbank outputs used directly as features without a DCT step, have become especially prominent with the rise of Deep Neural Network (DNN) acoustic models (Dahl et al., 2011; Mohamed et al., 2012). These features have shown to be effective for ASR tasks (Gao et al., 2015). However, their design is optimized for recognition rather than discrimination. They focus on mapping acoustically varying realizations of the same phoneme to a stable representation, which may cause them to overlook subtle differences that are crucial for detecting near-miss or borderline pronunciation errors.

As research progressed, it became increasingly clear that segmental acoustic features alone are insufficient to capture the full range of L2 deviations, leading researchers to include suprasegmental information. Non-native speech often deviates from native norms not only in phoneme accuracy, but also in prosodic patterns such as stress, rhythm, and intonation. Numerous studies have emphasized the importance of prosodic cues for assessment (Levis, 2005; Field, 2005). For instance, temporal measures like speech rate and articulation rate serve as strong indicators of fluency (Cucchiarini et al., 2002; Toivola et al., 2009), while pitch, intensity, and duration are useful for evaluating lexical stress, prominence, and intonation (Höhle &

Weissenborn, 2003; Shi et al., 2010; Ren et al., 2004; Van Santen et al., 2009). These suprasegmental features have demonstrated substantial contributions to intelligibility in L2 speech (Chun & Levis, 2020; Jackson & O'Brien, 2010). The wide range of potentially relevant features has prompted the proposal of standardized feature sets, most notably the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and its extended version, eGeMAPS (Eyben et al., 2015). These standard sets are typically extracted via toolkits such as openSMILE (Eyben et al., 2010). These resources provide a consistent, theory-driven set of measures covering acoustic, prosodic, and voice quality aspects (e.g., jitter, shimmer, harmonics-to-noise ratio). Despite these rich feature sets, a fundamental research gap persists. There is a lack of systematic, large-scale studies to identify which features most effectively distinguish native speech from non-native speech. Determining the most salient features for this purpose remains a key area for ongoing research.

Along with the acoustic features and suprasegmental cues discussed earlier, another promising approach is the use of articulatory features (AFs). Compared to acoustic features and suprasegmental cues, AFs provide unique benefits for detecting mispronunciations. While acoustic and suprasegmental features mainly describe the speech signal or prosody, AFs offer a structured view of speech based on articulatory dimensions, such as voicing, place, and manner of articulation (Browman & Goldstein, 1992; Stevens, 2000). This articulatory approach offers several advantages for identifying mispronunciations. It can give more interpretable and detailed feedback than simply labeling a segment as "incorrect" by pinpointing the specific articulatory aspect causing the error (Chen et al., 2022; Duan et al., 2017; Mao et al., 2018). Additionally, AFs are more robust to common distortions in L2 speech, since deviations can be modeled as partial changes along relevant articulatory dimensions-a level of flexibility usually missing in discrete phoneme systems (Benway et al., 2023; Li et al., 2016b). Therefore, AFs can supplement acoustic and suprasegmental features by offering a more detailed, linguistically grounded perspective for assessing pronunciation quality.

However, directly measuring articulation with invasive methods such as electromagnetic articulography (EMA) or real-time MRI is impractical for large-scale or classroom settings, especially in CALL or pronunciation training contexts. To address this, acoustic-to-articulatory inversion (AAI) models have been developed to estimate AFs from speech acoustics. Early studies demonstrated the feasibility of subject-independent AAI-based AF estimation (Ghosh & Narayanan, 2011; Kirchhoff et al., 2002; Toutios & Margaritis, 2003), and recent progress using DNN and self-supervised learning has further boosted inversion accuracy and generalization

(Chung & Kang et al., 2024; Liu et al., 2015; Xie et al., 2016). It is important to note, however, that the majority of this research has been conducted on native speech, relying on articulatory corpora recorded from native speech. Despite these advancements, significant challenges remain, particularly when extending this approach to L2 speech. The mapping from acoustic signals to articulatory space is inherently nonlinear and underdetermined, making AF estimation sensitive to speaker variability, limited training data, and recording conditions (Siriwardena et al., 2022; Wang et al., 2022b; Wieling et al., 2017). Studies also indicate that articulatory kinematics (such as tongue trajectories) do not always stay consistently linked with acoustic outputs under variations in speech rate or loudness (Mefferd & Green, 2010). Consequently, while AAI-derived AFs are more robust than systems based on discrete phonemes in handling partial or distorted articulatory deviations, their practical application in large-scale, cross-speaker, or E2E ASR-based L2 pronunciation assessment is still limited. Addressing these gaps calls for more systematic research, especially on how AFs, which are often derived from native-trained models, can best support both accurate mispronunciation detection and detailed, linguistically informed feedback within modern E2E assessment systems.

1.2.3 The paradigm shift towards End-to-End architectures

The third significant area in automatic pronunciation assessment involves the development of the underlying modeling architecture. The effectiveness of such a system relies on how effectively it captures the complex, nonlinear relationship between the acoustic signals and the intended phonetic units. Over the past two decades, this component has experienced significant changes, evolving from modular ASR pipelines to integrated end-to-end frameworks.

For many years, the dominant framework for automatic pronunciation assessment was the hybrid Gaussian Mixture Model (GMM)-Hidden Markov Model (HMM) pipeline, in which GMMs estimate frame-level acoustic likelihoods and HMMs model temporal phonetic sequences for forced alignment and scoring (Witt & Young, 2000; Young et al., 2002). Concretely, GMMs represent the distribution of acoustic feature vectors (e.g., MFCCs) for each sub-phonetic state as mixtures of Gaussians, while HMMs impose a sequential structure over these states and enable Viterbi-based forced alignment of phonetic units to the waveform. With the deep learning revolution, DNN-based acoustic models largely replaced GMMs in the early 2010s, yielding more discriminative acoustic posteriors and improved phoneme discrimination and GOP estimation (Hinton et al., 2012; Mohamed et al., 2012; Hu et al., 2015b). However, the modular pipeline, most notably its dependence on forced alignment and fixed phonetic units, constrains its ability to represent the variability

and ambiguity of L2 speech. Forced-alignment errors can propagate to scoring and are especially problematic for accented, child, or otherwise non-canonical speech (Mathad et al., 2021; Mahr et al., 2021). Moreover, the pipeline requires extensive phonetic transcriptions and curated mispronunciation labels for effective supervised training, which limits scalability in real-world CALL deployments.

Driven by the desire to simplify multi-stage pipelines and reduce error propagation, the ASR community has gradually moved toward E2E architectures. Early efforts on labeling unsegmented sequences, notably Connectionist Temporal Classification (CTC), provide a principled way to train sequence models without frame-level alignments (Graves et al., 2006). Subsequently, large-scale neural E2E systems such as DeepSpeech demonstrated that neural networks can replace multiple specialized modules and learn task-relevant representations directly from raw (or minimally processed) acoustic inputs (Hannun et al., 2014). Attention-based sequence-to-sequence models (e.g., Chorowski et al., 2015; Chan et al., 2016) and hybrid CTC/attention frameworks (Watanabe et al., 2017) further expanded the range of alignment-free modeling methods, while Transformer and Conformer architectures enhanced the modeling of long-range and local dependencies and raised ASR performance standards (Vaswani et al., 2017; Gulati et al., 2020).

The rise of E2E architectures has paralleled the integration of self-supervised learning (SSL) and large-scale pretraining, which is particularly impactful for L2 pronunciation assessment where annotated data are limited. Models like Wav2Vec 2.0 learn robust acoustic representations by masking and predicting latent speech units and can be fine-tuned with relatively small amounts of labeled data to achieve strong downstream performance (Baevski et al., 2020). Cross-lingual SSL variants, such as Self-Supervised Cross-Lingual Speech Representation Learning at Scale (XLS-R), extend this concept to multilingual pre-training and have been shown to enhance low-resource recognition, enabling better sharing of representations across languages (Babu et al., 2021; Conneau et al., 2020). Recent applied research demonstrates that Wav2Vec-style pretrained encoders significantly benefit mispronunciation detection and intelligibility assessment when fine-tuned on limited datasets (Xu et al., 2021; Yang et al., 2022; Shekar et al., 2023). These developments have prompted a shift away from heavily engineered, alignment-dependent pipelines toward more data-driven methods that involve pretraining and fine-tuning for pronunciation assessment.

Despite the advances outlined above, adapting E2E and SSL models for the specific diagnostic needs of MDD remains a challenging area of ongoing research. Many E2E architectures (e.g., CTC or attention-based sequence models) are primarily

optimized to reduce recognition errors and thus learn to produce the most probable phoneme or word sequence. They are not inherently designed to reveal detailed, sub-segmental phonetic deviations necessary for diagnostic feedback. As a result, the canonical ASR output space does not naturally provide access to articulatory or feature-level information (e.g., partial voicing, place or manner deviations) that is pedagogically valuable (Kheir et al., 2023; Zhang et al., 2020). Several recent studies therefore suggest redesigning or augmenting E2E models so they generate richer, more interpretable outputs for MDD—such as by integrating auxiliary articulatory or phonetic targets, multi-task objectives, or adopting architectures that explicitly model sub-segmental representations (Wu et al., 2021; Yan et al., 2020; Zhang et al., 2022). Representations like AFs are a promising source of complementary information, as discussed in Section 1.2.2, but it remains an open question how best to incorporate AFs into E2E encoders and alignment-free decoders to simultaneously improve detection accuracy and provide actionable, linguistically-informed feedback for L2 learners.

1.3 Research questions and outline

The literature reviewed in Sections 1.1 and 1.2 highlights the multifaceted nature of L2 speech assessment, encompassing both high-level functional communication and low-level signal characteristics. However, this review also reveals specific areas that warrant further investigation, particularly concerning the integration of objective acoustic measures with human-centered assessments and the technical constraints of modeling non-native speech. Consequently, this thesis does not aim to cover the entire spectrum of CALL challenges but focuses on a set of interrelated problems that are critical for advancing the accuracy and robustness of automatic systems. The following research questions are designed to systematically explore these areas:

RQ1: How can acoustic-phonetic features be leveraged to distinguish native from non-native speech? (**Chapter 2**)

RQ2: How do acoustic-phonetic features correlate with and predict the intelligibility of non-native speech? (**Chapter 3**)

RQ3: How do cumulated speech resources influence ASR performance and mispronunciation detection in pluricentric non-native speech? (**Chapter 4**)

RQ4: How can articulatory features be integrated to enhance mispronunciation detection and diagnosis in non-native speech? (**Chapter 5**)

The thesis is structured as a series of interconnected studies that follow a research trajectory moving from foundational signal analysis to complex diagnostic modeling.

Starting with a foundational investigation into the signal characteristics of nonnative speech (RQ1), **Chapter 2** presents a systematic investigation aimed at identifying the distinctive characteristics that effectively distinguish non-native speech from native speech. Unlike previous studies that often relied on specific or limited feature sets, this chapter adopts a data-driven approach, utilizing a comprehensive collection of standardized acoustic and temporal features. By employing automated feature selection techniques, the study aims to objectively identify the most discriminative indicators that separate the speech of L2 learners from native speakers. The identification of these critical features offers a useful reference for future research to focus on the most effective predictors, thereby facilitating the development of more efficient and accurate assessment models.

Building upon this foundational characterization, the research then scales to the functional level to examine the extent to which these features can predict speech intelligibility (RQ2). **Chapter 3** examines the assessment of Speech Intelligibility (SI) through two complementary studies, involving both subjective measurement and objective prediction. The first study aims to establish a reliable ground truth for assessment. It empirically investigates the relationship between two distinct subjective measures: Visual Analogue Scale (VAS) ratings and measures automatically derived from Orthographic Transcriptions (OTs). The study explores the reliability of these measures and the correlations between listener ratings and transcription-based accuracy scores. Importantly, this investigation is conducted at multiple levels of granularity, ranging from the sentence level down to word and subword levels (grapheme and phoneme), to examine how measurement reliability and correlations vary across different linguistic units. Building on this analysis, the second study of this chapter assesses the feasibility of objective assessment by investigating whether the features automatically extracted from the audio signal can be used to accurately predict human ratings through automated models. This progression from validating subjective metrics to developing objective models provides a strong framework for automatic intelligibility assessment.

To ensure that such assessment models are applicable in real-world linguistic contexts, the research then shifts to address the challenges of data variability and

resource scarcity (RQ3). **Chapter 4** investigates the effectiveness of using speech resources from a dominant language variety (Netherlandic Dutch) to enhance ASR models for a non-dominant variety (Flemish Dutch). This study focuses on the dual challenge of improving general recognition performance while ensuring accurate Pronunciation Error Detection. By exploring this cross-variety augmentation, the chapter aims to determine whether shared resources can effectively mitigate data scarcity issues for ASR, while also maintaining the model's ability to reliably detect pronunciation errors.

The final stage of this trajectory moves beyond global assessment towards providing precise and actionable diagnostic feedback (RQ4). **Chapter 5** introduces and assesses innovative modeling approaches to improve Mispronunciation Detection and Diagnosis in End-to-End systems. This chapter includes two related studies that utilize both customized architectures and fine-tuned pre-trained models. The first study systematically investigates methods to integrate inferred articulatory information into these End-to-End frameworks. It examines how this integration affects both segmental and sub-segmental output formats to measure improvements in detection accuracy and diagnostic precision. Building on these findings, the second study performs a comprehensive, multi-dimensional error analysis of these articulatory-enhanced models. It evaluates how the models behave across various factors such as utterance length and speaker variability to identify performance bottlenecks and to provide a deeper understanding of the balance between detection performance and diagnostic feedback.

Finally, **Chapter 6** summarizes the findings in the preceding chapters to provide comprehensive answers to the research questions. It discusses the theoretical and practical implications of the results for the field of CALL. The limitations of the research conducted are acknowledged, and these acknowledgements are used to propose promising directions for future research.


Hello
你好

Chapter 2

Distinctive Features for Classifying Spoken Native Versus Non-native Speech

This chapter is based on the following publication:
Wei, X., Cucchiarini, C., van Hout, R., & Strik, H. (2023). Distinctive Features for Classifying Spoken Native Versus Non-Native Speech. *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 51–55.
https://doi.org/10.21437/slate.2023-11.

Current applications of Automatic Speech Recognition (ASR) based technology for second language learning often require comparisons of native and non-native speech for evaluation and feedback purposes, which are generally based on limited sets of features that might not be the most optimal ones to characterize native as opposed to non-native speech. In the present study, we conducted a systematic comparison based on a large number of standardized acoustic and temporal features. The main aim was to gain insights into which features are most distinguishing. In turn, this knowledge can be employed to develop classifiers that are more suitable to evaluate non-native speech and to provide a solid basis for delivering feedback aimed at improving speech production. The findings indicate that most of the investigated features are significant and the temporal features are also distinctive. We discuss these results in relation to previous research and outline avenues for future investigations.

2.1 Introduction

In recent years, the demand for second language learning has increased rapidly. It is more and more important for people to effectively acquire one or more foreign languages. Due to the interference from the native language, it is difficult for non-native speakers to acquire native-like performance (Derakhshan & Karimi, 2015). Generally, the pronunciation errors made by non-native speakers are distortion errors, rather than absolute phone substitutions (Yoon et al., 2010). It is well known that pronunciation teaching requires intensive practice and feedback, which is limited by the shortage of qualified teachers. With the advancement of ASR technology, many ASR-based Computer Assisted Language Learning (CALL) applications have made outstanding contributions to developing programs that help improve the pronunciation quality of non-native speakers (Bai et al., 2020; Tafazoli et al., 2019). One of the aims of a CALL system is to pinpoint pronunciation errors and provide effective feedback. However, the shortage of non-native speech resources makes it challenging to build robust ASR-CALL systems. Many efforts have been made to overcome the lack of speech resources and improve the performance of non-native ASR (Duan et al., 2017; Duan et al., 2019; Zhang et al., 2020).

In the field of non-native ASR, one of the central research issues is to select informative features from speech signals. In Gao et al. (2015), researchers investigated how different acoustic features, PLP, MFCC, FBANK, and their combinations affect the DNN-based ASR model. Articulatory features treated as a representation of the movement of the pronunciation organs have also been widely used as features to detect pronunciation errors (Mao et al., 2018; Wei et al., 2017). Spectral features

and some temporal features have also been taken into account (Burgos et al., 2014). eGeMAPS (Eyben et al., 2015) is a set of 88 basic standard acoustic features for automatic speech analysis in various areas. In Baird et al. (2021), the eGeMAPS feature sets were also introduced for automatic speech emotion recognition. In an earlier study, researchers showed that a set of 103 features, including the eGeMAPS features and other 15-dimension features extracted through Praat, was useful for selecting the most distinguishing features for Dutch Chronic Obstructive Pulmonary Disease (COPD) patients in stable and exacerbated conditions, and in healthy conditions (van Bemmel et al., 2021). Additionally, in a non-native speech intelligibility study, researchers identified systematic differences in vowel duration between native and non-native speakers (Bent et al., 2008).

In an attempt to automatically evaluate the fluency of Dutch non-native speakers in read speech, researchers (Cucchiarini et al., 2000; Cucchiarini et al., 2002) extracted a wide range of temporal features, including speech rate, articulation rate, phonation ratio, etc., from native and non-native speech resources. Compared with ratings by human experts, speech rate obtained the highest correlation, while the articulation rate also played an important role in determining reading fluency. It is worth noting that the non-native speech rate was found to be more variable than the native speech rate, which means that both mean differences and variability in the signal should be taken into account (Baese-Berk & Morrill, 2015). In Toivola et al. (2009), researchers emphasized that native Finnish speakers were systematically faster than non-native speakers with regard to articulation rate. In addition, a study on the influence of ideology on the perception of non-native speech volume found that the participants perceived native-accented English as louder than non-native English (Crabtree, 2021).

Despite all the efforts that have been made, only a limited set of features are generally included in comparing native and non-native speech and little is known about which features are most characteristic and optimal for characterizing these types of speech. Especially, the pronunciation deviations of non-native speakers may be affected by age, accent, proficiency level, and the most characteristic features may vary in different languages or datasets, which makes this a more challenging task.

The central research question in the current study is which features are characteristic for differentiating between Dutch native and non-native speech. Based on the literature, we formulate a more specific hypothesis: non-native speakers display lower articulation rates and tend to speak with a lower volume due to a general lack of confidence in producing speech (Crabtree, 2021; Reves & Medgyes, 1994; Toivola et al., 2009).

2.2 Method

2.2.1 Speech material

The native read speech was extracted from "The spoken Dutch Corpus" (Corpus Gesproken Nederlands, CGN) (Oostdijk, 2002), which consists of 9 million words read by adults in the Netherlands and Flanders and enriched with various annotation layers. Two non-native read speech corpora were included, the first one is a Dutch non-native learning project named Idiomatic Second Language Acquisition (ISLA) in Hubers et al. (2021). As a Dutch idiomatic expression database, ISLA only includes idioms collected from various sources, without other kinds of formulaic expressions. Except for 374 idiomatic expressions from 394 native speakers, ISLA also contains a subset of 110 and 60 idiomatic expressions from German and Arabic Dutch L2 learners, respectively. In the present research, we selected the German and Arabic non-native read speech from ISLA. Meanwhile, a comparable amount of native read speech material as the German non-native read speech was selected from the CGN corpus. Another non-native speech material is a Dutch phonetic-rich speech corpus named JASMIN (Cucchiarini et al., 2008) that contains both Dutch native and non-native speech recorded by different age groups varying from children to elders. In the present study, we selected only read speech from forty-five adult non-native speakers, 17 males and 28 females, with various language backgrounds, such as Morocco, Turkey, etc. A more detailed description of the selected data is shown in Table 2.1.

Table 2.1: *Overview of speech data.*

Dataset	Num. of words		Num. of speakers
	Original	After outlier removal	
CGN	26758	25760	25
ISLA_German	26445	25379	41
ISLA_Arabic	7531	7165	23
JASMIN	42267	32345	45

2.2.2 Acoustic feature extraction and selection

The entire procedure of the proposed method is shown in Figure 2.1. By default, the openSMILE toolkit extracts the eGeMAPS feature sets over the entire audio file. The recordings comprise numerous utterances. Therefore, to automatically extract word-level features, a Forced Aligner (CLST; web version, https://webservices.cls.ru.nl/forcedalignment2; this program can align utterances to both the word and phoneme levels) was employed to generate word-level segmentations. Due to some incorrect boundaries taking place mainly at the beginning and end of the recordings, an outlier

detection was performed to automatically remove words with incorrect boundaries. The total numbers of words before and after outlier detection are shown in Table 2.1.

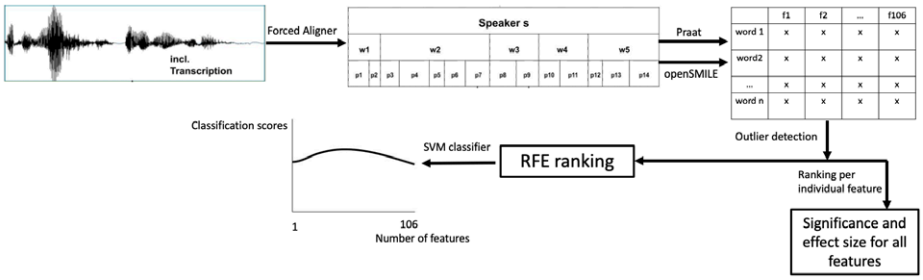

Figure 2.1: *A visualization of the automatic word-level feature selection.*

Subsequently, 103 word-level features were automatically extracted via Praat (Boersma, 2011) and openSMILE (Eyben et al., 2010). The 15 Praat features include duration, four formants, gravity centre, pitch variance, and the minimum, maximum, mean, and standard deviation of intensity and pitch. The 88 eGeMAPS feature set was extracted using openSMILE. In addition, three measures of articulation rate were calculated by dividing the numbers of syllables, graphemes, and phonemes by the duration of the word. Hence, a total of 106 word-level features were extracted for each recording.

The feature selection method we applied next was a combination of a Recursive Feature Elimination (RFE) and Support Vector Machine (SVM) classifier with a linear kernel, initially proposed by Guyon et al. (2002) and widely used in both bioinformatics and other research areas. This can be implemented by the 'scikit-learn' package (Pedregosa et al., 2011). Moreover, Leave-One-Subject-Out (LOSO; Sakar et al., 2013) cross-validation was implemented within classification to avoid overfitting and to prevent identity confounding. Each time the LOSO cross-validation selects one feature as a test set for iteration. The SVM classifier is used in both the RFE ranking and the follow-up evaluation of the selected features.

Finally, statistical analyses were performed for the ranked features. Welch's t-test was carried out for finding the significant feature sets, and the effect size (Cohen's *d*) for all the features was calculated. To examine distinctions between native and non-native datasets, the standardized difference between two means, which represents the effect size (Cohen's *d*), was calculated for all the features:

$$Cohen's\ d = \frac{M_1 - M_2}{S_{pooled}} \tag{2.1}$$

where M_1 and M_2 denotes the mean of the two groups being compared, while S_{pooled} is the pooled standard deviations for the two groups.

2.2.3 Evaluation metrics

The results of the SVM-based classifier, in general, can be divided into four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Therefore, Accuracy (ACC) was calculated to evaluate the performance, as shown in Eq. (2.2):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

Considering the uneven sample size among different datasets, we used another measure for binary classifier evaluation, Matthews Correlation Coefficient (MCC) as shown in Eq (2.3):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{2.3}$$

2.3 Results and discussions

In terms of the comparison between CGN and ISLA_German, in total, 94 of the 106 features obtained a p value smaller than 0.05 (Welch's t-test), of which the p value of 89 features was smaller than 0.001. This reveals that the majority of the chosen features are significantly different. As for the comparison between CGN and ISLA_Arabic, 91 features were found significant ($p < 0.05$), of which 87 features obtained a p value smaller than 0.001. For the binary comparison between CGN and JASMIN, 90 features were found significant, of which 84 features have p values less than 0.001. The effect size was also calculated for those significant features. Shown in Table 2.2 are the top-10 features ranked by RFE and corresponding Cohen's d rankings for the three pairwise comparisons. The first rank column in the table represents the ranking by RFE (r.RFE), while the second column indicates the ranking by Cohen's d (r.Cd).

Table 2.2: *At the top, the top 10 features by CGN vs. ISLA_German; at the middle, the top 10 features by CGN vs. ISLA_Arabic.at the bottom, the top 10 features by CGN vs. JASMIN. r.RFE= ranked by RFE; r.Cd = ranked by Cohen's distance; d = Cohen's distance; M = Mean; SD = standard deviation; sma3 = symmetric moving average filter for 3 frames long; sma3nz = non-zero conditional sma3; amean = arithmetic mean; stddevNorm = coefficient of variation.*

Comparison	Rank		*d*	Feature	M/SD	
	r.RFE	*r.Cd*			Native	Non-native
CGN vs. ISLA_German	1	3	1.742	intensity_mean	69.117/4.226	58.280/7.732
	2	19	0.853	HNRdBACF_sma3nz_amean	5.665/2.792	8.179/3.100
	3	12	1.174	spectralFlux_sma3_amean	0.298/0.144	0.120/0.158
	4	8	1.439	slopeV0-500_sma3nz_amean	0.033/0.020	0.003/0.021
	5	1	1.842	loudness_sma3_percentile80.0	0.960/0.353	0.388/0.260
	6	57	0.179	mfcc2V_sma3nz_amean	9.287/10.914	7.489/9.116
	7	2	1.808	loudness_sma3_amean	0.659/0.241	0.278/0.176
	8	27	0.578	loudness_sma3_stddevNorm	0.500/0.178	0.406/0.148
	9	5	1.732	equivalentSoundLevel_dBp	-24.994/4.241	-35.747/7.707
	10	9	1.313	slopeUV0-500_sma3nz_amean	0.003/0.020	-0.024/0.021
CGN vs. ISLA_Arabic	1	1	1.982	intensity_mean	69.117/4.226	57.718/9.365
	2	6	1.512	slopeV0-500_sma3nz_amean	0.033/0.020	0.001/0.024
	3	48	0.270	HNRdBACF_sma3nz_amean	5.665/2.792	6.454/3.348
	4	13	1.023	spectralFlux_sma3_amean	0.298/0.144	0.134/0.208
	5	4	1.593	loudness_sma3_amean	0.659/0.241	0.286/0.211
	6	60	0.187	mfcc2V_sma3nz_amean	9.287/10.914	7.305/9.310
	7	100	0.008	mfcc2_sma3_stddevNorm	-34.221/5474.65	4.074/238.54
	8	78	0.061	slopeV500-1500_sma3nz_amean	-0.012/0.013	-0.011/0.012
	9	42	0.362	F1frequency_sma3nz_amean	619.76/151.74	562.14/182.82
	10	57	0.201	F2frequency_sma3nz_amean	1546.8/278.20	1485.65/384.12

Table 2.2: *Continued*

Comparison	Rank		*d*	Feature	M/SD	
	r.RFE	*r.Cd*			Native	Non-native
	1	65	0.139	F3amplitudeLogRelF0_sma3nz_amean	-97.533/41.10	-92.624/29.948
	2	67	0.126	F2amplitudeLogRelF0_sma3nz_amean	-95.048/42.381	-90.504/30.393
	3	7	0.772	F2frequency_sma3nz_amean	1546.8/278.195	1708.335/130.445
	4	1	1.291	articulation_rate_grapheme	16.604/5.70	10.547/3.697
CGN vs. JASMIN	5	39	0.376	mfcc2_sma3_amean	9.601/8.782	12.851/8.530
	6	62	0.154	mfcc2V_sma3nz_amean	9.287/10.914	10.955/10.788
	7	16	0.625	alphaRatioV_sma3nz_amean	-12.519/6.794	-16.332/5.483
	8	12	0.678	slopeV0-500_sma3nz_amean	0.033/0.020	0.019/0.020
	9	63	0.151	F1amplitudeLogRelF0_sma3nz_amean	-91.354/45.832	-85.440/32.650
	10	27	0.491	F0semitoneFrom27.5Hz_sma3nz_amean	29.009/6.990	31.917/4.920

Of all the 94 significant features ($p < 0.05$) investigated for CGN versus ISLA_German, 19 features achieved a very large effect size ($d > 0.8$), the following 11 features still have a large effect size ($0.5 < d < 0.8$), and the last 40 features appear to have a small effect size ($d < 0.2$). As for the binary comparison between CGN and ISLA_Arabic, 16 of the 91 significant features reached a very large effect size, and a very small effect size was observed for the last 34 features. Regarding the comparison between CGN and JASMIN, 6 features achieved a significantly large effect size, 18 features with a large effect size, and the effect sizes of 32 features were small.

As we can see in Table 2.2, some features ranked high by both RFE and the effect size, especially for the binary comparison between CGN and ISLA_German. However, it is also clear that Cohen's distances for part of the top-10 features ranked by RFE are relatively smaller, especially for CGN versus ISLA_Arabic. For instance, the feature 'mfcc2_sma3_stddevNorm' ranked seventh by RFE but had a very small effect size. This discrepancy between the two rankings can be explained by the different aims of the two ranking approaches. The advantage of RFE is that the dependencies between features are taken into account, which aims to maximize the distinction between the two types of speech taking into account the potential redundancies between the

measures. Table 2.2 reveals a clear observation: when comparing CGN with the two datasets from the ISLA corpus, it becomes evident that 6 out of the top-10 features ranked by RFE are identical, whereas the count for the top-10 features ranked by Cohen's d is 9. In contrast, we also found only two features figured among the three comparisons, which is acceptable since ISLA_German and ISLA_Arabic have a similar recording environment and speech materials, while JASMIN is more phonetically rich and more diverse as to the backgrounds of the non-native speakers. This result provided an answer to the sub-question we formulated.

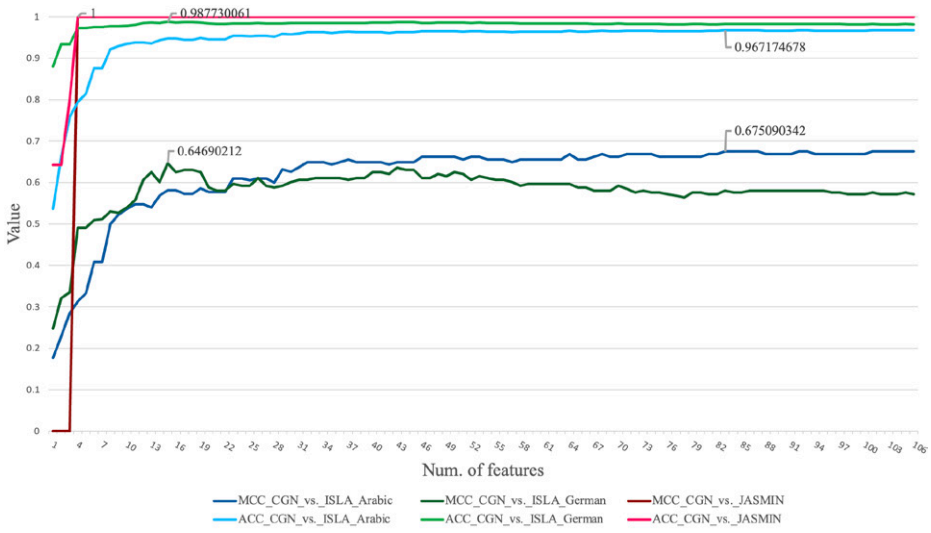

Figure 2.2: *The MCC and ACC values per number of RFE ranked features by SVM-based classifiers.*

In Table 2.2, the top 10 feature lists include some features related to loudness and features highly correlated with loudness like 'intensity', when comparing CGN with the two ISLA datasets, for which the values for non-native speakers are smaller than those of native speakers. This finding broadly supports previous research that, when speaking a second language, the volume of non-native speakers generally is lower than native speakers due to the lack of confidence (Toivola et al., 2009). Although a similar finding was not in the comparison between CGN and JASMIN, we found the feature 'articulation_rate_grapheme' ranked high by both RFE and Cohen's *d*. In addition, the other two articulation rate features, though ranked not too high by RFE, also ranked in the top 3 by Cohen's *d*. This result is in line with previous research that articulation rate shows distinctive to distinguish native from non-native speech.

The SVM-based classifiers adopted reflect how well the recordings can be separated by the chosen features. Figure 2.2 indicates the MCC and ACC curves per number

of features by RFE ranking. The highest values were achieved with the top 15 and 83 features for ISLA_German (MCC = 0.647; ACC = 0.988) and ISLA_Arabic (MCC = 0.675; ACC = 0.967), respectively. The different results, especially the number of features, may be due to the smaller size of the Arabic non-native speech. As for the comparison between CGN and JASMIN, it reached the highest MCC (=1) and ACC (=1) with only the top 4 features. As we can see in Figure 2.2, both curves for CGN versus JASMIN reached 1 sharply which is different from the other two models. This may be due to both CGN and JASMIN being phonetic-rich datasets, and the comparable balanced datasets (compared to the ISLA datasets) can make the convergence of the classification model faster.

2.4 Conclusions

To investigate the most characteristic features distinguishing Dutch native and non-native speakers, we applied an automatic feature extraction and selection approach. A large number of acoustic and temporal features were considered on the word level, including features extracted from Praat and openSMILE, as well as articulation rates. Furthermore, two different non-native read speech corpora were taken into account. RFE was implemented to select the set of most characteristic features and to determine their rank order in characteristic value within the combined set of features. In addition, the rank order of the individual features was computed on the basis of their effect size (Cohen's *d*). SVM-based classifiers with the LOSO scheme were also introduced to distinguish native and non-native speech according to the RFE ranking results.

For all three comparisons, at least 90 features achieved a *p*-value smaller than 0.05, which clearly showed that most of the explored features were significantly different between native and non-native speech. This observation answers the central research question about which features are characteristic for classifying Dutch native and non-native speech. With regard to the binary comparison between CGN and the two datasets from the ISLA corpus, close and moderate outcomes of the SVM classifier were observed. The best classification result was attained between CGN and JASMIN using only the top 4 RFE-ranked features, probably due to substantial dependencies between the features. We calculated the correlation between the features and indeed observed high correlations (up to 0.99).

All three articulation rate measures are in the top 10 of the features with the highest Cohen's *d*, only one of them (articulation_rate_grapheme) shows up in the

top-10 RFE list; also, here probably because of the high correlations between these three articulation rate measures. Moreover, we found that the distributions of all articulation rates are significantly lower for non-natives. In addition, many features related to loudness were observed for comparison between CGN and the ISLA datasets. These findings match our hypothesis that non-native speakers tend to have a lower volume and articulation rate, compared with native speakers.

Future endeavors could involve expanding the speech material and exploring alternative classification approaches. In addition, incorporating additional temporal features, for instance, rate of speech, phonation ratio, and mean length of runs, which have appeared to vary considerably between native and non-native speakers, should be considered (Cucchiarini et al., 2000; Cucchiarini et al., 2002).


Hello
你好

Chapter 3

Measuring and Predicting Speech Intelligibility in Non-native Speech

This chapter is based on the following publications:

Wei, X., Hout, R. van, Catia Cucchiarini, Reuvekamp, D., & Helmer Strik. (2023). Assessing Intelligibility in Non-native Speech: Comparing Measures Obtained at Different Levels. *Interspeech 2023*, pp. 964–968. https://doi.org/10.21437/interspeech.2023-1635.

Wei, X., Cucchiarini, C., van Hout, R., & Strik, H. (2023). Measuring Intelligibility in Non-native Speech: The Usability of Automatically Extracted Acoustic-Phonetic Features. *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 121–125. https:// doi.org/10.21437/slate.2023-23.

3.1 Assessing Intelligibility in Non-native Speech: Comparing Measures Obtained at Different Levels

Speech intelligibility (SI) is essential in communication and second language learning. In this study, non-native SI was measured through Visual Analogue Scale (VAS) scores and Orthographic Transcriptions (OTs) of read aloud sentences. Seven measures automatically derived from the OTs at word and subword levels were studied. The reliability of the intelligibility measures and the correlations between VAS scores and OT-based measures were also explored. Despite the different speaker language backgrounds, the recruited raters exhibited high scoring reliability. The correlations between VAS scores and OT-based measures were weak, corroborating previous assumptions that they refer to two related but distinct notions, comprehensibility (VAS) and intelligibility (OT). OT-based measures are reliable and valid indicators of SI. The results are discussed in relation to previous studies and avenues for future research are proposed.

3.1.1 Introduction

The growing application of Automatic Speech Recognition (ASR) based technology in the field of second language (L2) learning, can potentially alleviate the shortage of qualified teachers. Although L2 learners can practice their pronunciation with the help of such applications, it remains a big challenge to establish to what extent their speech is intelligible or comprehensible. In L2 learning, speech intelligibility (SI) generally refers to the extent to which listeners can actually understand L2 sentences, while comprehensibility refers to the difficulty or ease listeners experience in understanding utterances (Kennedy & Trofimovich, 2008). In speech pathology, on the other hand, slightly different definitions are employed (Yorkston et al., 1996). Both intelligibility and comprehensibility are affected by linguistic background, proficiency levels, speech rate, and several other factors (Munro et al., 2006). In L2 learning research researchers suggest that comprehensibility should be measured by collecting scalar judgments, e.g., through a Likert scale (Yorkston & Beukelman, 1978), or the Visual Analogue Scale (VAS; Finizia et al., 1998). Intelligibility, on the other hand, should be measured through manual orthographic transcriptions (OTs) of speech, as mentioned in Munro & Derwing (2015). However, various methods are employed to measure L2 intelligibility (Thomson, 2017), and in speech pathology both metrics are employed to measure intelligibility (Xue et al., 2020). For manual transcriptions, various speech materials can be employed, including whole sentences, isolated words or pseudowords, as well as Semantically Unpredictable Sentences (SUS) as is often done in speech pathology research. All these types of speech materials have their own advantages and disadvantages. In the case of

isolated words, the influence of the context on the listeners can be minimised. With the help of a speech recognition-based pronunciation trainer, in Burleson (2007), the intelligibility of English segments produced in isolated words by Chinese speakers was significantly improved even without teacher interaction. On the other hand, an obvious flaw of isolated words is their unnatural context, which makes it unclear how some specific phonemes in isolated words relate to SI in a more natural context. For these reasons, it seems preferable to transcribe whole sentences rather than isolated words. Sentences may be more intelligible because they contain context, this is a more natural condition than isolated, decontextualized words. In Xue et al. (2020), researchers designed three experiments with different speech materials including sentences and isolated words. The results showed that for all measures the intelligibility of sentences was noticeably higher than that of isolated words.

The above-mentioned methods rely on subjective human judgments, which are costly, time-consuming, and difficult to implement on a large scale. In addition, the assessment may be affected by the type of measurement or the listeners. For instance, the measurements collected from experienced listeners were less varied than those by inexperienced listeners (Mencke et al., 1983). Therefore, researchers explored some relatively objective approaches, such as the usability of speech technology to evaluate intelligibility, like ASR-based algorithms or ASR-free features, to predict SI scores (Middag et al., 2011; Spille et al., 2018). In Yu et al. (2015), a Bidirectional Long Short-Term Memory (BLSTM) and Multilayer Perceptron (MLP) - Linear Regression (LR) jointed model was proposed to automatically grade non-native speech.

The mentioned approaches may provide objective results comparable to those obtained with human ratings, which could alleviate the need to rely on human efforts. A limitation of these methods is the lack of adequate speech resources that are well transcribed. For this reason, researchers in the field of speech pathology have proposed a semi-automatic approach to evaluating the intelligibility of disordered speech (Ganzeboom et al., 2016). In this study, the intelligibility of dysarthric speakers measured by OTs was evaluated at sentence, word, and subword levels. The word and subword level ratings were automatically derived by forced alignment and conversion methods. The results proved that the introduced approach was feasible and reliable while providing a more detailed and informative measurement of intelligibility.

In the field of second language learning, there is no consensus on which method should be used to measure SI (Kang et al., 2018). In Xue et al. (2020), researchers compared five different methods to evaluate the intelligibility of six English distinct varieties. The listeners were required to assess the speech by responding to true or false statements,

scalar ratings, perception of nonsense and filtered sentences, and transcription of speech. The results showed that all the methods were effective for intelligibility measurement, but the correlations between the methods were not strong.

The studies discussed above were mainly in the field of pathological speech or L2 English, while relatively few studies have studied SI measurement for other languages. In the present research, an online listening experiment was conducted in order to investigate the SI of non-native speech as measured by VAS scores and measures automatically derived from OTs. The following research questions were addressed: 1) to what extent can VAS scores and OT-based measures provide a reliable basis for the assessment of non-native SI? 2) How strong are the correlations between VAS scores and OT-based measures?

3.1.2 Method

3.1.2.1 Speech material

The material was selected from a corpus named JASMIN (Cucchiarini et al., 2008), which includes both native and non-native speech from various groups: native primary school children, native secondary school students, non-native children, non-native adults, and native senior citizens. For each group around ninety-five hours of speech recordings were collected, half of which are read speech. Approximately two-thirds of the speakers in this corpus were recruited in the Netherlands and the rest in Flanders.

For non-native read speech, the corpus includes recordings from forty-six speakers. The reading material consists of phonetically rich sentences, stories, and general texts. Given that the raters had to both score and transcribe the recordings sentence by sentence, the listening materials could not be too long to prevent the raters from making 'mistakes' due to lack of memory instead of lack of intelligibility of the speakers. In order to give away as little context as possible to the raters, phonetically rich utterances were selected from the non-native speech part of the corpus. Unlike stories and general texts, the phonetically rich utterances are independent of each other.

In order to control the duration of the experiment and avoid deviations caused by fatigue, the same five sentences from nine non-native speakers were selected (see Table 3.1.1). Their recordings lasted on average 7 seconds. Additionally, loudness normalisation was applied for all the recordings to make sure these would be perceived as equally loud by the raters and for a more pleasant listening experience.

Table 3.1.1: *Sentences selected from the nine speakers.*

Id	Sentence
1	ik wou al om half drie hier zijn om alles in de etalage te zetten
2	de voetballer belooft zijn contractuele verplichtingen na te komen
3	de juffrouw rust een middagje uit en doet een dutje
4	de chauffeur tracht met wilde bewegingen de kuilen in de weg te omzeilen
5	de huiseigenaar kwam aan de deur om de huur op te halen

3.1.2.2 Rating procedure

Twenty-one expert raters were recruited through a call in a Facebook group that was not accessible to everyone. None of them was familiar with the materials used in the present research. Fourteen raters were certified L2 teachers, but all twenty-one had at least one year of experience in language teaching. Three are men and eighteen are women. On average, the age of the raters and years of experience as an L2 teacher are around 47 and 9.8, respectively.

The listening experiment was conducted using the program Radboud Online Linguistic Experiment Generator (ROLEG), which is a software application from Radboud University that can be used to create behavioural experiments and surveys. Before the listening experiment, the raters could familiarise themselves with the task by listening to extra recordings and receiving extensive instructions on the orthographic transcription and the VAS assessments.

The entire experiment, aside from intermission, lasted approximately 40 to 50 minutes and consisted of two parts with an option to pause in between. In the first part, the first three utterances in Table 3.1.1 were scored and transcribed by the raters. The last two sentences were played in the second part. In both parts, the recordings were played randomly. This way the raters had the option to pause and get a new impulse by hearing new sentences in the second part of the experiment.

Although the raters could replay the recordings multiple times, they were asked to repeat a recording no more than once (i.e., listen maximally twice). The raters first made the OTs and then assigned the VAS scores without OTs prompts. The OTs were without punctuation and could be grammatically or semantically incorrect and even include nonsense words since the raters were asked to transcribe precisely what they heard. One rater missed three recordings belonging to two speakers. Consequently, the total number of VAS scores and orthographic transcriptions is 942 (i.e., 9 speakers 5 sentences $\times$ 21 raters - 3 missed sentences).

3.1.2.3 Intelligibility measures

3.1.2.3.1 Sentence level measure

For VAS rating, a continuous bar was shown on the screen, where the left end denotes completely incomprehensible speech (score = 0) and the right end means perfectly intelligible speech (score = 100). Raters were encouraged to use the entire scale.

3.1.2.3.2 Word level measure

Orthographic transcriptions were compared with the reference transcriptions (the prompts in Table 3.1.1) after removing punctuation marks, and then Levenshtein distance was applied to calculate word accuracy (W_{Acc}) as shown:

$$Acc = 100 \times (N_{total} - N_d - N_s)/N_{total} \tag{3.1.1}$$

where N_{total}, N_d, and N_s represents the number of total words, deletion, and substitution, respectively.

3.1.2.3.3 Subword level measures

Grapheme strings were automatically converted to phoneme strings through an online Grapheme to Phoneme (G2P) tool (https://webservices.cls.ru.nl/g2pservice), and all subsequent subword level analyses were carried out for both the grapheme and the phoneme strings. The transcriptions of the spoken sentences were first aligned with the reference transcriptions by means of the 'ADAPT' algorithm (Elffers et al., 2005). For graphemes and phonemes, accuracy (G_{Acc} *and* P_{Acc}) was then calculated using equation (3.1.1), while distance (*Dist*) and the number of changes (*Ch*) using equations (3.1.2) and (3.1.3), respectively.

$$Dist = N_d \times C_d + N_i \times C_i + N_s \times C_s \tag{3.1.2}$$

$$Ch = N_d + N_i + N_s \tag{3.1.3}$$

where N_i represents the number of insertions. C_d, C_i, and C_s denote the cost for deletion, insertion and substitution. For grapheme, C_d and C_i *equalled* 1, and C_s was 2. As for phoneme, C_d and C_i were 3, and C_s was calculated by employing metrics with articulatory features. These different weights were employed in different versions of the 'ADAPT' algorithm that was used for alignment.

3.1.3 Results

3.1.3.1 Reliability of intelligibility measures

Boxplots of VAS scores and W_{Acc} are shown in Figure 3.1.1. The variance in W_{Acc} is less than in VAS, which may have consequences for their correlation with the OT measures.

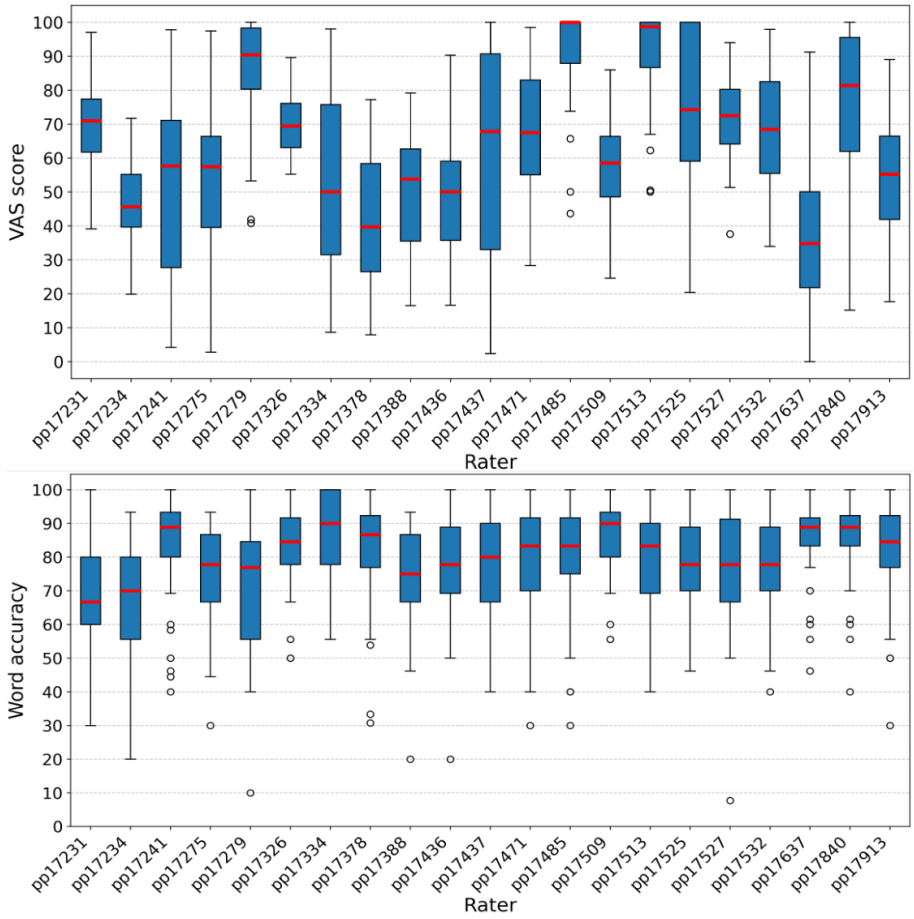

Figure 3.1.1: *Boxplots of the VAS scores (upper) and W_{Acc} (lower) per rater. The horizontal axes denote the twenty-one raters.*

In order to study the reliability, the Intraclass Correlation Coefficient (ICC) per sentence for VAS scores and the three Acc measures derived from the OTs were calculated, see Table 3.1.2. The ICC results were computed through SPSS (Cronk, 2016), and the parameter for model and type is 'Two-Way random' and 'consistency', respectively. It is apparent that the raters have a higher ICC for sentences 2, 3, and 5,

with the lowest ICCs for sentence 4. The majority of the coefficients are higher than 0.9, which reveals that the consistency in ratings among the raters is high for VAS and the three accuracy measures.

Table 3.1.2: *ICC for the intelligibility measures per sentence.*

Sentence	VAS	W_{Acc}	G_{Acc}	P_{Acc}
1	0.912	0.885	0.807	0.813
2	0.953	0.941	0.971	0.965
3	0.964	0.972	0.960	0.937
4	0.917	0.939	0.778	0.763
5	0.958	0.934	0.977	0.966

3.1.3.2 Statistical analysis of intelligibility measures

The mean, standard deviation (SD), median, and variance of all intelligibility measures are given in Table 3.1.3. For the sentence level, there are VAS scores, while all the Acc measures relate to lower levels (W, G, P). It can be observed in Table 3.1.3 that the mean values of G_{Acc} and P_{Acc} are similar, and both are larger than W_{Acc}, while the magnitude of the VAS scores is much lower. In addition, the number of changes at the grapheme level is larger than that at the phoneme level. Since all the Acc measures were derived from OTs, the correlations between W_{Acc} and VAS scores were further explored in 3.1.3.3. A closer inspection of the scattergram in Figure 3.1.2 indicates that the same W_{Acc} corresponds to various VAS scores, especially when W_{Acc} scores are between 70 and 90.

Table 3.1.3: *Mean, standard deviation, median, and variance of measures at different levels.*

Level	Measure	Mean	SD	Median	Variance
Sentence	*VAS*	62.99	24.28	64.06	589.74
Word	W_{Acc}	79.05	15.92	83.33	253.36
	G_{Acc}	92.62	8.09	95.0	65.48
Grapheme	G_{Dist}	8.53	7.13	7.0	50.84
	G_{Ch}	6.90	6.34	5.0	40.26
	P_{Acc}	92.37	7.73	94.23	59.82
Phoneme	P_{Dist}	17.52	15.99	14.0	255.69
	P_{Ch}	5.96	5.45	5.0	29.71

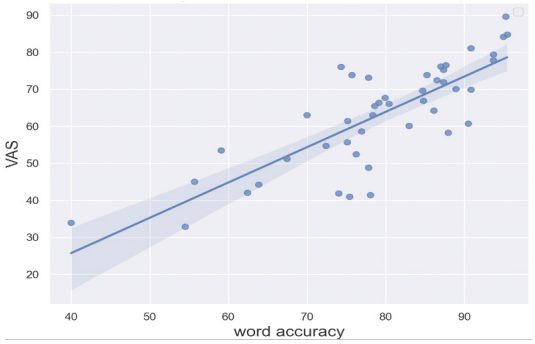

Figure 3.1.2: *Scattergram of W_{Acc} versus VAS at the sentence level. The shaded zone denotes the 95% confidence interval.*

In order to explore which specific words were relatively more challenging in terms of intelligibility, the OTs at the grapheme level were further analyzed by computing the correctness of the transcription of the content words. Table 3.1.4 presents the top 10 highest and lowest correctly transcribed content words. The top 10 highest correct words consist mainly of short and high-frequency verbs, while the top 10 lowest correct words are mainly long words or short words with the rounded front phonemes /y/ ('uu') or /œy/ ('ui'). For instance, the word 'huur' was transcribed in various ways, such as 'hu' (final r deletion), 'guur' (fricative instead of the semivowel), and 'heur' (a lower rounded vowel).

Table 3.1.4: *The top 10 highest and lowest corrected transcribed words with percentage.*

Highest correctness		**Lowest correctness**	
Word	**Percentage**	**Word**	**Percentage**
zijn	96.8%	contractuele	19.1%
kwam	92.0%	huiseigenaar	28.3%
halen	89.8%	kuilen	37.4%
zetten	87.8%	uit	43.9%
komen	87.2%	omzeilen	46.0%
wou	86.8%	bewegingen	49.2%
etalage	84.7%	middagje	52.4%
weg	84.0%	huur	57.2%
alles	83.1%	verplichtingen	58.2%
hier	81.5%	voetballer	59.6%

3.1.3.3 Correlations between intelligibility measures

The correlation coefficients among the explored intelligibility measures in Table 3.1.5 reveal that W_{Acc} is strongly correlated with the Acc measures at the subword level, while this is not the case for the VAS scores. The highest correlation obtained, as expected, was between G_{Acc} and P_{Acc} (= 0.919). It is important to note that the correlation coefficient between the VAS and W_{Acc} is weak (= 0.325; see Figure 3.1.2), and even lower with the two Acc measures at the subword level (see Table 3.1.5).

Table 3.1.5: *Correlations between measures at word and subword levels.*

	Grapheme			**Phoneme**		
	Acc	*Dist*	*Ch*	*Acc*	*Dist*	*Ch*
VAS	0.296	-0.302	-0.249	0.299	-0.222	-0.229
W_{Acc}	0.751	-0.647	-0.573	0.693	-0.451	-0.463

3.1.4 Discussion and conclusions

In the present study, a listening experiment was conducted aimed at a comprehensive assessment of non-native SI. Speech materials from nine non-native speakers selected from a speech corpus were assessed by native speakers with experience in language teaching through VAS scores and OTs.

Table 3.1.2 shows that most of the ICC values per sentence are higher than 0.9, indicating a strong reliability of the raters' measurements. Sentence 4 was more difficult to understand, which might be caused by the length and/or the number of long words (sentence 4 contained three long words listed with the lowest correctness in Table 3.1.4).

Three accuracy measures were automatically derived from OTs and prompts at the word, grapheme, and phoneme levels, as well as distance scores and the number of changes. The results for all eight measures shown in Table 3.1.3 indicate that all VAS mean values are lower than those of W_{Acc}. In addition, Figure 3.1.2 shows that the same W_{Acc} may correspond to various VAS scores, while the values of G_{Acc} and P_{Acc} were close and significantly better than the above two SI measures, which is in line with outcomes in Xue et al. (2020) and Ganzeboom et al. (2016). This is understandable since only one grapheme or phoneme transcription error will result in a score of 0 for the whole word. Additionally, one phoneme may relate to more than one grapheme and then its correctness depends on the associated grapheme to a certain degree (Abur et al., 2019; Ganzeboom et al., 2016).

As to the comparison between VAS and of W_{Acc} our results are in line with those of previous research employing these measures for dysarthric speech (Ganzeboom et al.

2016), in which VAS scores appeared to be lower than W_{Acc}, indicating that raters tend to judge sentences to be less intelligible than they in fact are. However, a different trend was observed in (Xue et al., 2020), which may be ascribed to differences between speech materials and human raters. Our outcomes are also more in line with those of research on non-native SI by Kang et al. (2017). In this study, five different methods of intelligibility measurement were investigated and found to be weakly correlated with each other. We also found a low correlation of VAS scores with the Acc measures, while the correlations between the three Acc measures were high, which is understandable as these were all derived from the OTs.

The above findings show that although VAS scores can be obtained more easily, they show more variation between raters. The advantage of OTs is that they make it possible to analyze intelligibility at the word and subword level, which is important in the field of L2 learning, but also in speech pathology. By calculating the seven measures used in the current paper, much more detailed information at various levels can be obtained, e.g., the easiest and most difficult words shown in Table 3.1.4.

To summarize, we can conclude that all the measures investigated were reliable for assessing different aspects of non-native SI. The selection of the most suitable measurement should be made in relation to the intended purpose. VAS scores appear to be more intuitive and would seem to be more appropriate for assessing the perceived level of speech intelligibility, which is often referred to as comprehensibility. OT measures, on the other hand, seem to be more direct measures of SI and seem appropriate to support research on specific pronunciation problems in a substantial way.

Future work could study the effect of specific words and sounds in natural sentences on the assessment of intelligibility. It is also important to investigate the possibility of automatically generating orthographic transcriptions and assessing intelligibility without human-made transcriptions, through ASR. The rapid developments in ASR technology suggest new avenues of research in this direction that are definitely worth exploring. Another option could be to investigate how SI measurement by experts, as in this study, relates to that by native raters without experience in language teaching. This would certainly be interesting in terms of ecological validity, as non-native speakers will not always be evaluated by language teachers, but eventually by the native speakers with which they will engage in conversation in their daily lives.

3.2 Measuring Intelligibility in Non-native Speech: The Usability of Automatically Extracted Acoustic-Phonetic Features

Speech intelligibility (SI) plays an important role in second language learning. It can be influenced by many factors and various approaches have been explored to measure it. In this study, the intelligibility of Dutch non-native speech was measured by Visual Analogue Scale (VAS) and word accuracy (W_{Acc}) that was automatically derived from orthographic transcriptions (OTs). A large number of acoustic-phonetic features were automatically extracted from the audio files. To deal with the multicollinearity issue, feature reduction through different regression approaches was investigated. The results revealed that, as a whole, the LASSO regression approach outperformed (highest R^2 at 0.52, lowest RMSE and MAE) the other explored regression methods. The regression models predict the intelligibility measures assigned by human raters well. The obtained findings indicate the usability of automatically extracted acoustic-phonetic features to provide a basis for the assessment of SI.

3.2.1 Introduction

Accurate and objective prediction of speech intelligibility (SI) is important for research and applications in many fields of speech communication (development of hearing aid devices, dysarthric speech diagnosis and therapy, second language learning) and many efforts have been made in this direction (Arai et al., 2020; Janbakhshi et al., 2019; Xue et al., 2021). SI is generally defined as the percentage of speech that a listener can understand, but definitions can vary according to the task. For instance, in telecommunication SI is an evaluation of the loss of the transmission channel.

In second language learning (SLL), SI refers to the actual understanding of utterances by listeners (Kennedy & Trofimovich, 2008). Other than in native speech, intelligibility in non-native speech is generally affected by the linguistic backgrounds and proficiency level of the non-native speakers (Munro et al., 2006), which makes it more challenging and crucial to measure SI in this context. One approach to measuring intelligibility, as mentioned by Munro & Derwing (2015), is to transcribe the speech by the listeners. In the field of pathological speech, another widely used method is to collect scalar judgments from human raters. These can be equal-appearing interval scales like the Likert scale (Yorkston & Beukelman, 1978), or ratings of intelligibility on a horizontal line with a point like Visual Analogue Scale (VAS) scores (Finizia & Dotevall, 1998). Since these evaluations can contain an element of subjectivity, they are generally collected from multiple listeners and averaged for further analysis.

All these types of manual measurements are highly costly and time-consuming. Consequently, researchers have been looking for alternatives to automatically obtain objective measures of intelligibility that do not need to rely completely on intensive human efforts. With the rapid improvements in Automatic Speech Recognition (ASR) technology, researchers have experimented with ASR-based algorithms and ASR-free features to predict SI (Middag et al., 2011; Spille et al., 2018). In spite of these efforts, little progress has been made in fields like SLL and speech pathology, which clearly suffer from resource scarcity (O'Brien et al., 2018). In addition, it is unclear how exactly ASR measures are related to SI. In Xue et al. (2021), with limited data, researchers studied the intelligibility of dysarthric speech through stepwise logistic regression to gain more insights into these relationships.

To a certain extent, the performance of the aforementioned approaches depends on the specific selection of features. Previous research has revealed that some acoustic-phonetic features are related to intelligibility, for instance, pitch (Levis, 2018; Koffi, 2021; van Maastricht et al., 2016), intensity (Koffi 2015; Koffi, 2021), and formant frequencies (Bohn & Flege, 1992; Zhi & Li, 2021). In Xue et al. (2019), researchers introduced the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature set (Eyben et al., 2015), which contains the above-mentioned three kinds of features, to predict phoneme level intelligibility. The eGeMAPS feature set, which was developed for speech analysis in various areas, contains a set of standardized features that were chosen according to their demonstrated theoretical relevance and their potential to distinguish important aspects of speech production. The eGeMAPS feature set can easily be extracted using openSMILE (Eyben et al., 2013), which calculates all parameters in a standardized way. Cucchiarini et al. (2002) extracted a set of temporal features including speech rate, articulation rate, and phonation time ratio from native and non-native speech and found the highest correlation with human scores for speech rate.

Although a wide range of features was taken into account, not all features are equally important. Especially for high dimensional data (Ayesha et al., 2020), some of the features are highly correlated, which may lead to multicollinearity issues and affect the results of the analysis. Hence, it is vital to remove the highly correlated, redundant features and retain as many relevant and informative features as possible. Fewzee & Karray (2012) compared Elastic net regression, greedy feature selection, and Principal Component Analysis (PCA) for dimension reduction in emotional speech recognition tasks. The Least Absolute Shrinkage and Selection Operator (LASSO) regression is another welcomed approach to selecting informative features. For instance, Zhou et al. (2010) proposed a LASSO regularizer-based framework that outperformed

other algorithms for multi- task feature selection. Another method with a specified threshold conducted with R was also recommended (Deepanshu, 2022).

In the current research, SI of non-native speech was explored as measured by VAS scores and OTs in relation to acoustic-phonetic features derived from the speech recordings. We study the combination of three dimension reduction approaches with a cross-validated stepwise linear regression to explore two research questions: 1) Which of these three approaches is more suitable for predicting SI scores? 2) To what extent can the acoustic-phonetic features be used to predict SI scores?

3.2.2 Method

3.2.2.1 Speech material

Forty-five utterances were selected from a speech corpus named JASMIN (Cucchiarini et al., 2008), containing the same five phonetically rich utterances for nine non-native speakers (5 female; mean age 31). In this way, we restricted the duration of the listening experiment and avoided errors caused by fatigue. We carefully examined potential deviating effects caused by specific utterances or speakers but did not find such effects. Additionally, loudness normalization was applied to remove loudness differences between the recordings.

3.2.2.2 Intelligibility measures

Twenty-one expert raters (18 female; mean age 47) engaged in the listening experiment. None were familiar with the materials used in the present research. Fourteen raters were certified L2 teachers, and the others had at least one year of experience in language teaching (mean 9.8 years).

The human raters could listen to the recordings multiple times but were instructed to repeat them maximally once (i.e., listen maximally twice). They assigned VAS scores ranging from 0 (completely incomprehensible) to 100 (comprehensible), as well as made accurate transcriptions of the utterances (OTs: Orthographic Transcriptions). They were required to transcribe as precisely as possible what they heard, including nonsense words and non-grammatical utterances. Word accuracy (W_{Acc}), a measure of the percentage of correctly transcribed words at the utterance level, was then calculated using the Levenshtein distance between the prompts and OTs, as shown in Eq. 3.2.1.

$$W_{Acc} = 100 \times (N_{total} - N_d - N_s)/N_{total} \tag{3.2.1}$$

where N_{total} , N_d , and N_s denotes the total number of words, deletions, and substitutions, respectively. Hence, for each transcription, two intelligibility measures were obtained: VAS score and W_{Acc}. In total, there are 945 combinations: 21 listeners times 45 utterances. However, one rater missed three utterances belonging to two speakers. As a consequence, the total number of OTs, W_{Acc} and VAS scores, is 942. The mean VAS score was 63.00 (median 64.06; minimum 0, maximum 100). The mean W_{Acc} was 94.49 (median 100; minimum 15.38, maximum 100).

The reliability between the raters was measured as Intraclass Correlation Coefficients (ICC), the mean and standard deviation of ICC are 0.94 and 0.02, 0.94 and 0.03 for VAS and W_{Acc}, respectively. A limited number of 5 utterances were selected, each one recurring 9 times in the listening experiment, which could potentially disturb the SI assessment, but it turned out that the reliability was high for both SI scores, even if the recordings were presented randomly. This indicates that the raters were able to assess the intelligibility of an utterance properly, even though they had already heard it. The correlation between VAS and W_{Acc} is weak (Pearson's correlation = 0.325), which is in line with previous research that VAS is more appropriate for measuring perceived speech comprehensibility, while OTs are more direct measures of SI (Munro & Derwing, 2015).

3.2.2.3 Acoustic-phonetic features

The acoustic-phonetic features consist of two groups of features derived for each recording. The first part contains 15 features that were extracted by Praat (Boersma, 2011), including the minimum, maximum, mean, and standard deviation of pitch and intensity, the mean of the absolute values of the pitch slope, formants 1 to 4, duration, and the centre of gravity. Another 88 eGeMAPS feature set was obtained by the default configuration in OpenSMILE. Speech rate, the number of words divided by duration, was added. Therefore, there are 104 features in total.

3.2.2.4 Features and statistical analysis

Multicollinearity is an important concern because of the extensive array of often related features. For dimension reduction, several regression methods were explored: LASSO regression (Tibshirani, 1996), Elastic-net regression (Zou & Hastie, 2005), and Redundant Variable Removement (RVR; Deepanshu, 2022). These three methods all result in smaller subsets of selected features. For all three methods, the selected features were then entered in a forward-backward stepwise linear regression (SLR) to predict the SI scores (VAS scores and W_{Acc}). Besides, the Leave One Out Cross-Validation (LOOCV) scheme was applied for the SLR model to prevent signaling spurious effects caused by dependencies on the utterance or speaker level. The Akaike Information Criterion (AIC; Sakamoto et al., 1986) was utilized in the stepwise procedure to decide

which acoustic-phonetic features had to be included in the final regression model. Results are regression equations for the two SI scores, the averaged values for R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and a final subset of features, which is a subset of the features selected by the regression algorithms.

The RVR method consists of the following steps: 1) generate absolute values of the correlation matrix for all the features; 2) replace all diagonal values in the matrix with NAs and compute the mean of each column; 3) calculate the rank of the mean values on descending order and reorder correlation matrix based on the rank; 4) check if the matrix(i, j) larger than the threshold, then calculate the mean value of the row of the matrix(i,) and matrix(-j,), remove variable i if matrix(i, j) has the highest mean absolute correlation value. After several rounds of iterations, the mean absolute correlation coefficients among all the features are smaller than a specified threshold. In the present study, the threshold we selected was 0.5.

Table 3.2.1: *Results for the two SI measures: VAS score and* W_{Acc} .

Method	VAS score					W_{Acc}				
	Selected features	**Final features**	**R^2**	**RMSE**	**MAE**	**Selected features**	**Final features**	**R^2**	**RMSE**	**MAE**
Elastic-net	8	4	0.33	11.74	9.25	4	3	0.20	10.81	8.15
LASSO	24	16	**0.52**	**10.33**	**7.61**	18	11	**0.48**	**8.66**	**6.87**
RVR-0.5	27	18	0.09	23.48	16.92	27	16	0.02	35.31	16.12

RVR-0.5 denotes the RVR method with a threshold of 0.5. Selected features denote the number of features selected by the regression methods; final features denote the number of features in the final model after applying SLR analysis in combination with the LOOCV scheme.

3.2.3 Results

3.2.3.1 Feature selection

The VIF scores of the entire feature set were all 'NaN', which indicated serious collinearity. After conducting feature selection, all the VIF scores were between 1 and 5. With regard to VAS prediction, the average VIF score for Elastic-net, LASSO, and RVR-0.5 was 1.16, 2.51, and 2.05, respectively. As for W_{Acc} prediction, the average VIF score for the three methods was 1.12, 1.49, and 1.65, respectively. It is apparent that the VIF scores revealed moderate correlations. In Table 3.2.1, the results for the three explored approaches are presented, the column 'Selected features' is the number of features after applying the approach in question, and the column 'Final features' is the number of features after applying stepwise linear regression in combination with cross-validation (LOOCV). It is clear that LASSO had the largest R^2 and the lowest RMSE and MAE values for both SI measures.

Table 3.2.2: *Summary of the final model in SLR for the intelligibility measures' prediction with selected features' coefficient, standard errors of the coefficients (Std. Err), the z value, the p-value, and the significance levels (Sig. level; 0.001 – '***'; 0.01 – '**'; 0.05 – '*'; 0.1 – '.'; 1 – ' ').*

	Feature	Coef.	Std. Err	z value	p-value	Sig. level
VAS	***F1bandwidth_sma3nz_stddevNorm***	-361.97	50.083	-7.2273	4.928e-13	***
	MeanVoicedSegmentLengthSec	-117.14	24.967	-4.692	0.0000027	***
	F3	0.012751	0.0003887	3.2807	0.0027742	**
	F1	-0.009389	0.002968	-3.1635	0.0037337	**
	shimmerLocaldB_sma3nz_stddevNorm	-30.337	10.552	-2.875	0.0040397	**
	logRelF0.H1-A3_sma3nz_amean	0.45237	0.15954	2.8356	0.0084019	**
	loudness_sma3_stddevFallingSlope	1.8563	0.66229	2.8029	0.0090923	**
	F3bandwidth_sma3nz_stddevNorm	60.754	21.842	2.7815	0.0095724	**
	slopeV0-500_sma3nz_stddevNorm	-0.10681	0.040351	-2.647	0.0131805	*
	duration	-0.001735	0.0007458	-2.3261	0.0274762	*
	F2bandwidth_sma3nz_stddevNorm	57.192	24.616	2.3234	0.0276422	*
	slopeUV0-500_sma3nz_amean	287.66	142.44	2.0195	0.0531037	.
	slopeV500-1500_sma3nz_stddevNorm	9.6094	5.0822	1.8908	0.0690398	.
	mfcc1_sma3_amean	0.75186	0.40713	1.8467	0.0753848	.
	FosemitoneFrom27.5hz_sma3nz_meanFallingSlope	-0.022163	0.013292	-1.6675	0.1065696	
	speech rate	-3.3309	2.9456	-1.1308	0.267733	
W_{Acc}	***MeanVoicedSegmentLengthSec***	-110.94	25.393	-4.3687	0.0001168	***
	logRelF0.H1-A3_sma3nz_amean	0.57644	0.14817	3.8904	0.0004588	***
	F1	-0.009548	0.0026769	-3.5669	0.0011288	**
	spectralFlux_sma3_amean	52.72	15.497	3.4019	0.0017690	**
	F3	0.0091701	0.0034219	2.6798	0.0114022	*
	mfcc1_sma3_amean	0.93491	0.35249	2.6523	0.0121923	*
	FosemitoneFrom27.5hz_sma3nz_stddevFallingSlope	0.015533	0.0058625	2.6495	0.0122752	*
	F3bandwidth_sma3nz_amean	-0.05095	0.019722	-2.5834	0.0144010	*
	shimmerLocaldB_sma3nz_stddevNorm	-25.923	10.984	-2.3601	0.0243364	*
	F3frequency_sma3nz_stddevNorm	-231.54	98.521	-2.3501	0.0248979	*
	F1bandwidth_sma3nz_stddevNorm	-112.18	53.368	-2.1021	0.043261	*

3.2.3.2 Prediction of intelligibility scores

Therefore, features selected by LASSO were utilized to fit the SLR models (LASSO-SLR). Table 3.2.2 presents results for the final LASSO-SLR models followed by the LOOCV scheme for the prediction of both SI scores. All the features in Table 3.2.2 are relevant according to the AIC criterion. The 7 features in bold overlapped within the two SLR final models. Positive coefficients denote positive relations to the criterion (the SI score), and negative coefficients show negative relations. Even though the significance levels vary, all of the 7 overlapping features have the same sign (positive or negative). Scatterplots of the predicted values versus the human ratings are shown in Figure 3.2.1. Most of the data points are around the regression line and the 95% confidence intervals (the shaded zone) for both predictions are small, particularly for the VAS prediction.

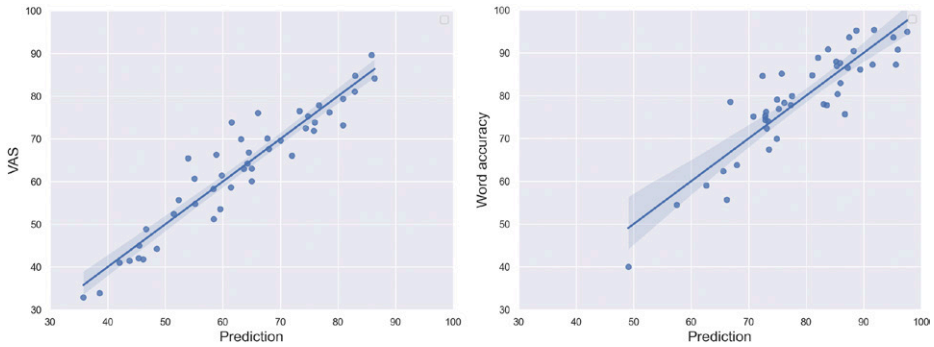

Figure 3.2.1: *Scattergrams of the two intelligibility ratings and predicted values. The vertical axis denotes the intelligibility rantings and the horizontal axis indicates the predicted values of the final LASSO-SLR models.*

3.2.4 Discussion and conclusions

The present research investigated whether and to what extent acoustic-phonetic measures of non-native SI, obtained by automatically processing audio recordings, can be used to estimate the intelligibility of non-native speech. For this purpose, their relation with two types of human ratings of SI was studied, i.e., VAS and W_{Acc}. We extracted a large number of acoustic-phonetic features. In total, 104 features: speech rate, 15 measures from Praat, and 88 OpenSMILE features.

To alleviate collinearity, feature selection was applied through three different regression approaches: Elastic-net, LASSO regression, and RVR. These approaches resulted in a linear regression model with selected feature sets to predict the SI scores (VAS and W_{Acc}). A model is better if R^2 is higher, and if RMSE and MAE are lower. For the RVR method, we explored a specified threshold at 0.5. An advantage of Elastic-net in comparison to RVR is that it employs fewer features, but it performs better for the predictions of both SI measures. LASSO regression retained more features and

showed better results than Elastic-net, which is in line with the results in Sridhara et al. (2020). The quantitative results for the three models presented in Table 3.2.1 show that LASSO selected as many features as possible while giving the best results: R^2 is the highest and RMSE and MAE are the lowest. These results answer our first research question.

In Table 3.2.2 the features and corresponding values in the final models are presented, together with their significance levels. In the final model for W_{Acc}, significant results were obtained for 11 of the 18 features, while in the final model for the VAS scores, this is the case for 16 of the 24 features. There is considerable overlap (7 features) in the features of the two models, and most of the coefficients of these features also have the same sign (positive or negative), indicating that those features have similar effects on the SI scores predicted by these models. Moreover, the common features were found to be significant in previous research, such as the F0 range was positively associated with sentence intelligibility (Bradlow et al., 1996), F1 carries 80% of the acoustic energy of vowels which benefits the intelligibility of vowels (Koffi, 2021), and F3 was found to be a significant feature when comparing the Japanese adults' and children's production of /l/ and /ɹ/ in American English (Aoyama et al., 2023).

Scatterplots of the predicted SI values versus the human ratings are presented in Figure 3.2.1, using the LASSO-SLR models with the LOOCV scheme. It can be observed that the majority of the data points are close to the regression lines and that the 95% confidence intervals (the shaded zones in Figure 3.2.1) are rather small, especially for VAS prediction. All these findings thus indicate that both VAS and W_{Acc} can be fairly well predicted with the final LASSO-SLR models through a limited number of automatically derived and selected acoustic-phonetic features, which answers the second research question. Similar outcomes have been obtained in the field of dysarthric speech (Xue et al., 2021).

It is interesting to observe that, compared with the model for W_{Acc} prediction, higher R^2 was obtained by the model for VAS prediction, revealing that the selected acoustic-phonetic features are better at fitting the model for the prediction of VAS scores. Similar outcomes were found in Xue et al. (2021), compared to W_{Acc}, a higher correlation between the computed acoustic-phonetic probability index and VAS was obtained. However, we also observed higher RMSE and MAE of the model for VAS prediction, implying comparable worse prediction quality. In other words, the selected feature can better fit the human scores but result in worse quality for VAS. This finding is in line with previous research (Munro & Derwing, 2015) that SI

is usually measured through OTs whereas what is measured through VAS score is perceived intelligibility often referred to as comprehensibility.

Since significant differences in speech rate have been observed between native and non-native speakers (Cucchiarini et al., 2002), it is surprising that the feature 'speech rate' was only present in the final model for the VAS score with a low significance, while it is not part of the final model for W_{Acc}. This unanticipated finding may be explained by the fact that although speech rate might be different between native and non-native speakers, it might not be an important factor in predicting SI scores within the group of non-natives only. This requires further research.

In summary, a wide range of acoustic-phonetic features were investigated to predict two SI scores. LASSO regression outperformed other explored approaches to reduce feature dimensions while retaining as many features as possible. Meanwhile, the SI scores were well predicted by the LASSO-SLR models. These findings provide insights into the automatic evaluation of intelligibility for low-resource second language learning. Further research could investigate other dimension reduction methods and the precise impact of the individual acoustic-phonetic features in more detail.


Hello
你好

Chapter 4

Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: Cumulating Speech Resources in a Pluricentric Language

This chapter is based on the following publication:
Wei, X., Cucchiarini, C., van Hout, R., & Strik, H. (2022). Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: cumulating speech resources in a pluricentric language. *Speech Communication*, *144*, 1-9.

The shortage of large-scale learners' speech corpora and precise manual annotations are two major challenges for automatic L2 speech recognition and error detection in L2 speech, especially for non-dominant varieties of pluricentric languages. In these cases, collecting and annotating large non-native (L2 learner) corpora for all language varieties is often unattainable. In this study, we investigated ways of addressing these problems through conventional and transfer learning Deep Neural Network (DNN) based Automatic Speech Recognition (ASR) and ASR-based pronunciation error detection (PED) by cumulating Netherlandic Dutch and Flemish Dutch speech resources. First, we show that for ASR the baseline system can be improved by combining the Netherlandic Dutch and Flemish Dutch datasets. Next, through the knowledge learned from models trained on the Netherlandic Dutch data, the Flemish Dutch learners' ASR model can be further improved. In order to evaluate the performance of the PED algorithms in the absence of learner speech data with pronunciation error annotations, we introduced plausible pronunciation errors in the native corpora based on knowledge from Flemish learner speech, in order to simulate non-native speech errors. For PED we found that the results are much better for a GOP classifier trained on Flemish Dutch data than for one trained on Netherlandic Dutch data. PED produced worse results when the Netherlandic Dutch data were merged with the Flemish Dutch data, while for ASR, lower WERs were attained. Whether adding Netherlandic Dutch data to Flemish Dutch data is beneficial, thus seems to depend on the specific task the data are used for. We discuss these results, compare them to those of related research and suggest avenues for future research.

4.1 Introduction

Pluricentric languages (PLCLs) are languages with minimally two language varieties belonging to different geographic centers that are commonly different nations. The varieties in question are often important symbols of the identity of the nations involved. The origin and rise of these varieties may have been influenced by a host of factors with historic, economic, and political origins (Clyne, 1997). In general, PLCLs have an official status in at least two nations. As of 2018, at least 43 languages are considered PLCLs (http://www.pluricentriclanguages.org/plc-languages).

Dutch is a West-Germanic language spoken by about 24 million people, most of whom live in the Netherlands (17 million), in Belgium (Flanders, Brussels; 6.5 million), and in Suriname (400, 000). According to the Taalunie, the Dutch language policy organization, Dutch is an official language in the Netherlands, Belgium, Suriname, and in the Caribbean islands of Aruba, Curaçao and Sint Maarten (https://taalunie.

org/informatie/112/taalunie-union-for-the-dutch-language). While Dutch (Northern) Dutch was seen as the same standard language variety in both the Netherlands and Flanders for many years, this situation started to change in the second half of the 20th century (Willemyns, 2013; Lybaert & Delarue, 2017). Nowadays, a distinction is made between the standard variety of the Netherlands (Netherlandic Dutch) and the standard variety of Flanders (Flemish Dutch). Dutch is now officially considered a pluricentric language and the recognition of both the Netherlands and Flanders as developing centers was carefully taken into account when setting up specific research programs funded by the Dutch and Flemish governments to stimulate the development of language and speech technology resources for the Dutch language as the Spoken Dutch Corpus (Oostdijk, 2002), the BLARK (Strik et al., 2002) and STEVIN (Spyns & Odijk, 2013). At present, both varieties of Dutch can be considered to be well-equipped in terms of language and speech technology resources, but there are domains in which paucity of resources still constitutes a problem (Odijk, 2012). This applies, in particular, to research on Automatic Speech Recognition (ASR) of L2 learners or non-native speech for which there are still limited resources, a situation that seems to apply to all languages other than English.

For another domain that is obviously affected by seriously limited resources, pathological speech, Yılmaz et al. (2016) investigated the benefits of combining speech data from different varieties of the same pluricentric language, Netherlandic and Flemish Dutch. They obtained promising results showing that ASR performance on Flemish Dutch pathological speech could be improved in this way. Against this background, it is interesting to study whether combining speech data of different varieties of the same pluricentric language also helps improve ASR performance in another domain that suffers from data sparseness, the speech of learners of Netherlandic and Flemish Dutch as a second language (L2 learners), often denoted as non-native speech.

Improving ASR technology for this type of speech can be relevant in several ways. First of all, ASR performance on non-native speech is usually lower, see 4.1.1, which hinders the development and use of ASR-based applications for L2 learners. Second, ASR technology can be employed to develop advanced Computer Assisted Language Learning (CALL) applications that can support L2 learners in acquiring oral skills, which usually receive less attention in L2 instruction because they require individual guidance and feedback. In this context, dedicated applications of ASR, such as Pronunciation Error Detection (PED), can be very effective, but, again, these require specific resources for training the PED algorithms, see 4.1.2. In this latter field of research, obtaining suitable data is even more problematic because the training

algorithms require learners' speech data with error transcriptions at the segmental level, which are difficult to obtain, thus leading to an even more severe lack of in-domain data.

In this paper, we investigate whether speech data of Netherlandic and Flemish Dutch can be usefully cumulated to improve ASR performance on non-native Flemish Dutch speech. Moreover, because these learners often need specific ASR-based CALL tools, we study whether combining speech resources can be used to the benefit of ASR-based Pronunciation Error Detection. As no non-native speech resources with an appropriate annotation of actual speech errors are available, we simulated pronunciation errors in our datasets.

We would like to emphasize that although some of the models and approaches adopted in this study have been used previously in studies addressing ASR and PED in English, it is still important and informative to investigate whether and to what extent these methods also work for other languages that are less resource-rich. This particularly applies to pluricentric languages, as we often see that especially the so-called "non-dominant varieties" do not possess sufficient amounts of speech data for specific applications.

We address the following two research questions in our study:

1) To what extent does including Netherlandic Dutch speech data (dominant in the amount of data) enhance ASR performance on non-native Flemish Dutch?
2) To what extent does including Netherlandic Dutch speech data (dominant in the amount of data) help improve Pronunciation Error Detection in non-native Flemish Dutch?

4.1.1 Non-native speech recognition

With the rapid advancement of automatic speech recognition (ASR), researchers have made considerable progress on non-standard speech tasks such as the recognition of regional accents and non-native speech (Biadsy et al., 2011; Shon et al., 2018; van Doremalen et al., 2010). Figure 4.1 illustrates a typical flowchart of an ASR system. The system includes four components. The first consists of processing the audio file for training and generating acoustic features, such as Mel-frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). The following module is the acoustic model (AM), which combines the knowledge of acoustics and phonetics, and generates AM scores for the feature sequences. The AM used in ASR-based systems has developed from the original template-based model (Hazen et al., 2009;

Lee & Glass, 2012) to a context-dependent Gaussian mixture hidden Markov model (Zheng et al., 2007). Owing to the development of deep learning and its applications in speech recognition, most current ASR systems are based on neural networks-HMM (Lo et al., 2020; Mao et al., 2018a). The third component, named language model (LM), estimates the probability of sequences by learning the relationship between words in the training text, also called LM scores. Finally, the decoder calculates the AM and LM scores for the given feature sequences and hypothetical word sequences. The word sequences with the highest overall scores are the output, the results in Figure 4.1.

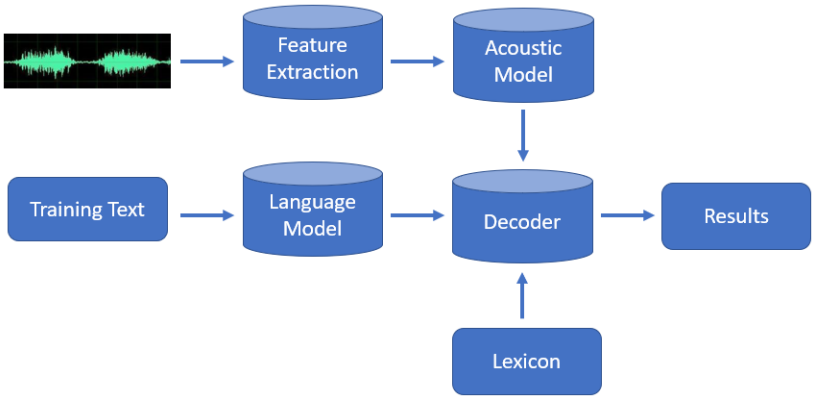

Figure 4.1: *Schematic diagram of a standard automatic speech recognition system.*

Once the input speech waveform is converted into a sequence of feature vectors O, the decoder tries to find the most probable word sequences W, as shown in the following equation:

$$\begin{aligned} \widehat{W} &= \underset{W}{argmax}\, P(W|O) = \underset{W}{argmax} \frac{P(O|W)P(W)}{P(O)} \\ &\cong \underset{W}{argmax}\, P(O|W)P(W) \end{aligned} \tag{4.1}$$

where $P(W)$ represents the LM probability, and the AM probability $P(O|W)$ can be calculated according to Eq. 4.2:

$$\begin{aligned} \mathrm{P(O|W)} &= \textstyle\sum_q P(O,q|W)P(q|W) \\ &\cong max\pi(q_0) \textstyle\prod_{t=1}^{T} a_{q_{t-1}-q_t} \prod_{t=0}^{T} P(q_t|O_t) \big/ P(q_t) \end{aligned} \tag{4.2}$$

The $P(q_t|O_t)$ is the state (senone) posterior probability estimated through the Neural Network (NN), $P(q_t)$ is the prior probability of each state obtained from the training data.

In general, speech recognition accuracy is lower for non-native speakers than for native speakers (Park & Culnan, 2019), which is mainly due to the mismatch in pronunciation between these two groups and the more considerable variability in

non-native speech. Non-native speakers often make specific pronunciation errors that are usually caused by their mother tongue. This is even more so for non-native speakers who are still in the process of learning a(n additional) language, often referred to as L2 learners. For instance, L2 learners of Netherlandic Dutch may have problems with the distinction between Dutch /i/ and /I/, because there is no such distinction in their mother language, as is the case for Japanese and Italian Dutch L2 learners (Truong et al., 2004). Moreover, L2 learners may produce non-existing words or words not present in the lexicon, which will also lead to lower recognition accuracy. AM, as one of the most important modules of an ASR-based system, has received the most attention in speech recognition. Various kinds of AMs can and have been explored to improve the performance of ASR systems. Directly using non-native speech data to train the acoustic models is the most straightforward approach to deal with the mismatch problem (Leung et al., 2019; Lin et al., 2020). However, this kind of data is less available and more difficult to collect than native speech data. Hence, several approaches have been proposed to tackle the issue of insufficient non-native speech data. For instance, multi-task learning is a widely used method for some related tasks of the same language (Duan et al., 2016; Mao et al., 2018b), such as recognizing adults' and children's speech by one model. The multitask-based NN model typically has more than two inputs and outputs but shared hidden layers, which means more data resources are available to train the NN. An alternative method is multi-lingual or cross-lingual acoustic modeling. Most of the early research based on multi-lingual learning was focused on the Tandem and Bottleneck features (Schultz & Waibel, 1998; Thomas et al., 2012), while the DNN-HMM based model did not receive more attention until recently (Duan et al., 2019; Joshi et al., 2015; Seide et al., 2011). Similar to the multi-task learning model, the multi-lingual learning model also has multiple inputs and outputs, as well as shared hidden layers. However, the inputs and outputs of multi-lingual learning normally come from different languages, such as Dutch and Mandarin Chinese. Transfer learning is another approach that is used extensively, especially for low-resource tasks (Nazir et al., 2019; Yan et al., 2020). Aimed to develop a better performing system for a new task quickly and effectively, this approach gains and maintains the knowledge learned from one or more similar tasks and applies it to new tasks.

4.1.2 Pronunciation error detection

ASR is aimed to convert an incoming speech signal into a text representing the words contained in that speech signal. Since words are pronounced variably or in a more reduced form, the ASR algorithms are trained in such a way that they can deal with variability and recognize variants of the words in question. In other words, ASR is a rather tolerant process. PED, on the other hand, aims at pinpointing incorrect

pronunciations, which means that it is a strict rather than a tolerant process. PED can be implemented at both segmental and suprasegmental levels (Meng, 2009). The segmental level mainly includes phones and words, while the suprasegmental level involves pitch accent (Ren et al., 2004), rhythm (Ito et al., 2006), and lexical stress (Höhle & Weissenborn, 2003). The pronunciations and canonical transcriptions of an L2 learner are firstly fed into the AM for forced alignment to calculate the confidence scores, which indicate the similarity between the non-native pronunciation and the canonical pronunciation. Various types of confidence measures have been investigated for PED, such as likelihood ratio (Franco et al., 1999) and posterior probability (Zhang et al., 2008). Owing to the development of DNN-based acoustic modeling, Goodness of Pronunciation (GOP) scores, a variation of posterior probability, have been widely used in current DNN-based models (Hu et al., 2015b). GOP scores can be derived from the AM trained on native or non-native speech data with pronunciation errors (Franco et al., 2010). The GOP score for a target phone p is computed as:

$$\mathrm{GOP}(p) \approx log \frac{P(p|o;t_s,t_e)}{max_{\{q\in Q\}}P(q|o;t_s,t_e)} \tag{4.3}$$

where p represents the canonical phone, q is the phone with the highest posterior probability, and Q indicates the set of all possible phones.

In addition, the log posterior of phone is computed as:

$$\log \mathrm{P}(p|o;t_s,t_e) \approx \frac{1}{t_e-t_s}\sum_{t_s}^{t_e} log \sum_{s\in p} P(s|o_t) \tag{4.4}$$

where t_s and t_e are the start and end indices of this phone, obtained by forced alignment; $\{s \in p\}$ is the set of context-dependent phones; o_t is the input feature at frame t; and $P(s|o_t)$ is the frame-level posterior score, which can be obtained by aligning the utterance with the corresponding canonical phone sequence.

For the sake of training a robust PED model, a large amount of learners' speech resources with error transcriptions are essential. This makes it harder to train PED models for low-resource pluricentric language varieties. Witt & Young (1997) proposed a GOP-based algorithm that was tested on artificial errors such as substituting /a/ for /i/ in a native corpus. A similar approach, but one that based its substitution strategy on actually observed L2 pronunciation errors, revealed that this procedure could provide satisfactory results (Kanters et al., 2009). In the case of Dutch L2 learning, mispronunciations often occur in vowels (Cucchiarini et al., 2009; Neri et al., 2006) since the Dutch have a rich inventory of 15 full vowels (van der Harst et al., 2014). For the

purpose of studying how PED in non-native Flemish Dutch speech can be improved, we adopted the previously used approach of artificially introducing well-attested pronunciation errors. Considering the non-native Flemish Dutch speech, in this case, is that of Francophone L1 speakers (see below), we introduced pronunciation errors made by Francophone learners of Dutch such as /ɑ/ vs. /a:/ and /I/ vs. /i/ (Hiligsmann & Rasier, 2006; see also van der Harst et al., 2014 for significant pronunciation distinctions in vowels between Netherlandic and Flemish Dutch, including /ɑ/ and /I/) to investigate to what extent our PED algorithms can detect these errors.

4.2 Methodology

As explained above, AM is an essential component in an ASR system. In this study, both conventional and transfer learning DNN models are explored and implemented. The Netherlandic Dutch and Flemish Dutch speech resources in different combinations were taken into account to train these models.

4.2.1 Speech data

Two corpora have been used in this study, named CGN (Oostdijk, 2002) and JASMIN (Cucchiarini et al., 2008). The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) is a corpus of standard Dutch as spoken by adults in the Netherlands and Flanders that is enriched with various annotation layers. This corpus contains about 9 million words. JASMIN is a corpus that includes both native and non-native speech recordings by different age groups: native primary children, native secondary students, non-native children, non-native adults, and native senior citizens. For each group, nearly 95 hours of speech were recorded, around half of which is read speech. The speakers were recruited in the Netherlands and Flanders. About two-thirds of the data was collected in the Netherlands, the rest in Flanders. The non-native Flemish speech data come from Francophone L1 speakers, while the non-native Dutch speech data consists of speakers with various L1s.

In this study, only part of the native and non-native adult read speech from CGN and JASMIN was extracted, respectively. The details of the four datasets are shown in Table 4.1. According to the Common European Framework of Reference for Languages (CEFR), the proficiency of the L2 learners in the JASMIN corpus can be classified at the A1, A2, and B1 levels. The number of L2 learners at the three proficiency levels, in the Netherlandic Dutch part, is 10, 25, and 10, respectively, while in the Flemish part of the corpus, there are 9, 9, and 10 learners, respectively. It can be observed that there are more native data than non-native data and that there are more females

than males. For non-native ASR modeling, we randomly and separately divided these four datasets into two subsets according to the utterances, 90% of which were used as the training set and 10% as the test set, as clarified in Table 4.1. Hence, there is no utterance overlap in each training set and corresponding test set, but there is speaker overlap. Therefore, we also implemented 10-fold cross-validation experiments aiming to verify whether the speaker overlap affects the performance of the proposed models. For each dataset, first, we randomly divided the entire dataset equally into 10 subsets according to the number of speakers. Each time we chose 9 subsets as the training set, and the remaining one as the test set. All the training procedures are the same as those of the other ASR models mentioned in section 4.2. Subsequently, the results after 10-time training were averaged and treated as the final result.

In the CGN and JASMIN corpora, and for the datasets we use, about two-thirds of the data are Netherlandic Dutch, and one-third is Flemish Dutch. Therefore, there is twice as much data for Netherlandic Dutch compared to Flemish Dutch, and thus Netherlandic Dutch is the variety for which more data is available.

Table 4.1: *Speech material. Rows 3 – 5: (3) Number of speakers (Total – Male & Female), (4-5) total number of utterances, and the duration in the training and test sets.*

	Native (CGN)		**Non-native (JASMIN)**	
	N1 Netherlandic Dutch	**F1 Flemish Dutch**	**N2 Netherlandic Dutch**	**F2 Flemish Dutch**
Speakers	324 – 144 M & 180 F	149 – 70 M & 79 F	45 – 17 M & 28 F	28 – 10 M & 18 F
Training set	30285 (25.96 hrs.)	15388 (14.46 hrs.)	2288 (3.77 hrs.)	1047 (1.43 hrs.)
Test set	3364 (2.91 hrs.)	1709 (1.57 hrs.)	572 (0.97 hrs.)	261 (0.36 hrs.)

4.2.2. Non-native speech recognition

All the acoustic models in this study were implemented in Kaldi (https://kaldi-asr.org/), an open-source toolkit widely used for speech recognition research. The training procedures of the models mentioned in the present study followed the standard Kaldi recipe. The acoustic feature we adopted is a 39-dim MFCC+Δ+ΔΔ vector. After forced alignments, context-dependent phone labels were used to train corresponding DNN acoustic models. We also introduced the i-vector as another input feature of DNNs since the age, nation, and gender of L2 speakers vary. Additionally, the Softmax function was applied at the output layer. The DNNs include 6 hidden layers with 2048 sigmoid nodes at each layer.

In the present study, we focus on acoustic modeling, with the aim of making better comparisons between different acoustic models possible. We use the same lexicon and language model for all our experiments. The lexicon contains all the words present in the datasets we use, and the tri-gram language model (LM) is trained on all the data we use by means of the SRI Language Modeling (SRILM) Toolkit (http://www.speech.sri.com/projects/srilm/).

4.2.2.1 Conventional DNN-based acoustic modeling

Inspired by the great improvements of DNN-based acoustic modeling, we adopt state-of-the-art DNN approaches to train different ASR models. For the names and datasets of the models, see Table 4.2. Since more Netherlandic Dutch speech resources are available, we first considered employing the Netherlandic Dutch native dataset for acoustic modeling (Con(ventional)-N1). As a comparison, an AM using the Flemish Dutch native dataset (Con-F1) was also trained. Subsequently, the Netherlandic Dutch and Flemish Dutch native datasets were combined as one training set for modeling (Con-N1-F1). In addition, we also explored the combination of the Netherlandic Dutch native and non-native datasets as the training set (Con-N1-N2) and the Flemish Dutch native and non-native datasets as the training set (Con-F1-F2), respectively. In order to evaluate the impact of the Netherlandic Dutch non-native speech data on the model, a model only using the Flemish Dutch non-native data (Con-F2) was investigated and then merged with non-native Netherlandic Dutch data (Con-N2-F2). Finally, all four aforementioned datasets were combined as the training set (Con-all).

Table 4.2: *Conventional DNN-based ASR models for non-native Flemish Dutch learners (F2). N = Netherlandic Dutch, F = Flemish Dutch; 1 = native speech, 2 = non-native speech.*

Model name	Training set	Test set	Pre-trained model
Con-N1	N1	F2	-
Con-F1	F1	F2	-
Con-N1-F1	N1 + F1	F2	-
Con-N1-N2	N1 + N2	F2	-
Con-F1-F2	F1 + F2	F2	-
Con-F2	F2	F2	-
Con-N2-F2	N2 + F2	F2	-
Con-all	N1 + N2 + F1 + F2	F2	-
TL-N1	F1	F2	Con-N1
TL-N1-N2	F1	F2	Con-N1-N2
Ref-N1	N1	N1	-
Ref-F1	F1	F1	-

4.2.2.2 Transfer learning-based acoustic modeling

Transfer learning can be traced back to 20 years ago and has been successfully applied in a wide range of research fields (Hajiramezanali et al., 2018; Kunze et al., 2017). This approach focuses on sorting knowledge learned from one task and applying it to related tasks. As a consequence, we investigated the possibility of transferring the knowledge gained from the Netherlandic AMs to non-native Flemish Dutch models. The schematic diagram of a DNN and transfer learning-based model is demonstrated in Figure 4.2. We assume that, in this study, the DNNs consist of shared hidden layers but task-dependent output layers. The shared hidden layers are used to transfer the knowledge of the AMs trained with the Netherlandic Dutch speech resources to our target models.

In transfer learning, pre-trained acoustic models are required, which in our case are trained on the Netherlandic Dutch L1 (TL-N1), and the combination of the Netherlandic Dutch L1 and L2 datasets (TL-N1-N2), see Table 4.2. Next, in both cases, the Flemish native data (F1) is used for transfer learning to retrain the Softmax layer of the pre-trained DNN model. The resulting two models are tested on the Flemish non-native data (F2), as shown in Table 4.2. In the first ten rows in Table 4.2, the test set is F2. In order to have reference values, we also trained two conventional DNN models named Ref-N1 and Ref-F1 with the Netherlandic Dutch native and Flemish Dutch native datasets, respectively. Their results can be found in the last two rows of Table 4.4.

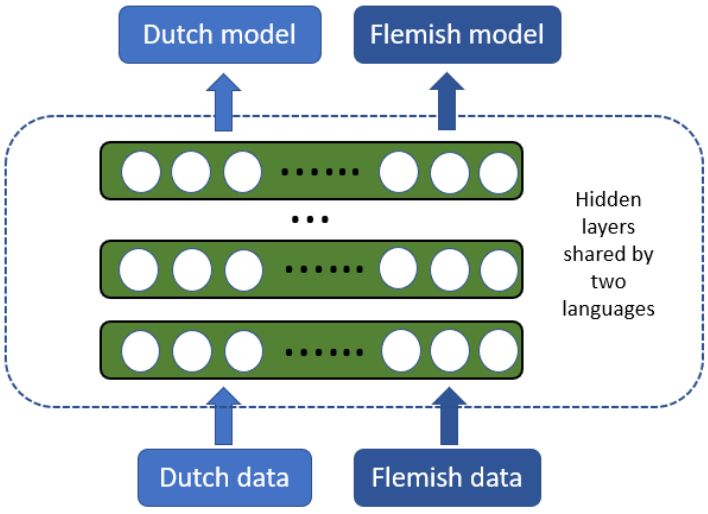

Figure 4.2: *Diagram of transfer learning-based acoustic modeling with shared hidden layers.*

4.2.2.3 Word error rate (WER)

WER, as a widely used measure to evaluate the accuracy of an ASR or machine translation system, is introduced to assess the performance of the AMs in this study. This function is derived from the Levenshtein distance and, by definition, it operates at the word level. In general, the output word sequence hypothesized by the ASR system is aligned with the reference word sequence. Therefore, WER can be calculated as:

$$WER = \frac{I + D + S}{N} \times 100 \tag{4.5}$$

where represents the number of insertions, is the number of deletions, is the number of substitutions, and is the total number of words in the reference word sequence.

4.2.3 GOP-based pronunciation error detection

The GOP algorithm was first proposed by Witt & Young (1997) and was later used and improved by many researchers. This algorithm aims to calculate the likelihood ratio between the realized phone and the canonical phone. It has been widely applied in computer-assisted pronunciation training (CAPT) to evaluate learners' pronunciation proficiency. Figure 4.3 demonstrates the procedure of the proposed DNN-GOP model. First, the acoustic features derived from the speech are fed into the AM. Subsequently, with phone level alignment information, the frame-level posteriors derived from the Softmax layer of DNNs are used to compute phone level posteriors. The GOP scores can be obtained by comparing the posteriors of the target phone with those of the most competing ones.

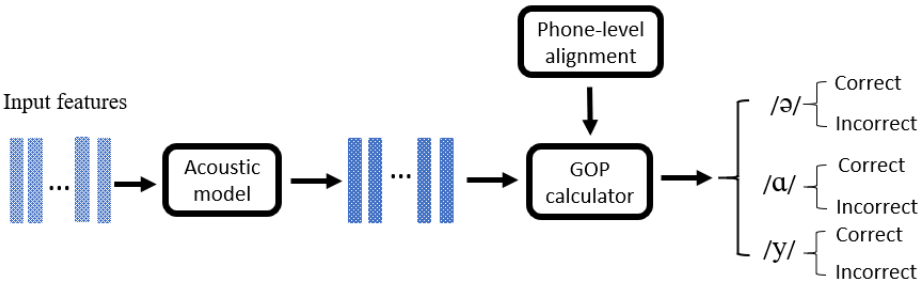

Figure 4.3: *Overview of proposed DNN-GOP model.*

In addition to GOP scores, a threshold is essential to determine whether each phone is correct or incorrect. In this study, we adopted a phone-independent threshold, obtained from the Equal Error Rate (EER), to determine whether the target phone is correct or incorrect. The results, in a PED task, are normally presented in four forms: correctly accepted (CA), i.e., correctly pronounced phones were detected as

correct; correctly rejected (CR), i.e., mispronunciations were detected as incorrect; false accepted (FA), i.e., mispronunciations were detected as correct; false rejected (FR), i.e., correct phones were detected as incorrect. Therefore, the False Acceptance Rate (FAR), False Rejection Rate (FRR), Scoring Accuracy (SA), and Cohen's kappa (κ) can be calculated as below:

$$FAR = \frac{FA}{FA + CR} \tag{4.6}$$

$$FRR = \frac{FR}{FR + CA} \tag{4.7}$$

$$SA = \frac{CA + CR}{CA + CR + FA + FR} \tag{4.8}$$

$$\kappa = \frac{2(CA * CR - FA * FR)}{(CA + FA)(CR + FA) + (CA + FR)(FR + CR)} \tag{4.9}$$

In addition, we adopt Recall, Precision, and the F1 score to evaluate the GOP-based models. These measures can be computed both for CA and CR as shown below:

$$Recall\ of\ CA = \frac{CA}{CA + FR} \tag{4.10}$$

$$Precision\ of\ CA = \frac{CA}{CA + FA} \tag{4.11}$$

$$Recall\ of\ CR = \frac{CR}{CR + FA} \tag{4.12}$$

$$Precision\ of\ CR = \frac{CR}{CR + FR} \tag{4.13}$$

$$F1\ score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4.14}$$

Generally, sufficient amounts of non-native speech material with corresponding transcriptions of pronunciation errors are not available. For this reason, alternative procedures have been developed to evaluate the performance of PED algorithms on the limited amount of data and annotations available. To evaluate the PED approaches in this study, three pairs of artificial errors were introduced: (1.) /ɑ/ - /a:/, (2.) /i/ - /a:/, and (3.) /i/ - /I/. All the artificial errors were introduced to the Netherlandic Dutch and Flemish Dutch native training sets and test sets independently before model training. For each pair, half of the occurrences in the native speech datasets were changed to their corresponding errors. The same three GOP-based models were trained for each pair of artificial errors, as described in Table 4.3. Similar to section 4.2.2.1, the first model of each set served as the baseline model. Obviously, there is no speaker overlap between the Netherlandic Dutch and Flemish Dutch native datasets.

Table 4.3: *The characteristics of the GOP-based models.*

Vowel pair	Model name	Training set	Test set
/ɑ/ - /aː/	GOP-1.N	N1	F1
	GOP-1.F	F1	F1
	GOP-1.N+F	N1 + F1	F1
/i/ - /aː/	GOP-2.N	N1	F1
	GOP-2.F	F1	F1
	GOP-2.N+F	N1 + F1	F1
/i/ - /I/	GOP-3.N	N1	F1
	GOP-3.F	F1	F1
	GOP-3.N+F	N1 + F1	F1

4.3 Results

In order to assess the effect of cumulating Netherlandic Dutch speech data in the Flemish Dutch learners' acoustic models, we conducted experiments on non-native ASR and PED and present the results in this section. Next, we present the results of our GOP-based PED systems.

4.3.1 ASR performance of different acoustic models

We assume that, in some cases, available data for a pluricentric language-based ASR system is insufficient. Hence, the baseline acoustic model is firstly trained with data from the variety with more speech resources, the Netherlandic Dutch L1 training dataset. In this part, we adopt WER to evaluate both the conventional DNN-based models and transfer learning-based models. In addition to the models mentioned above, we also provide models using only Netherlandic Dutch and Flemish Dutch native datasets to have reference values, as shown for the models Ref-N1 and Ref-F1. Table 4.4 describes the WERs of all the conventional DNN-based models, the transfer learning-based models, and the reference models.

From Table 4.4 we can see that the performance of the baseline model Con-N1 is unsatisfactory. Considering that the training set and test set of this model are from different varieties and that there is a mismatch between the native and non-native speech resources, 77.52% is a plausible result. Subsequently, when the training set was replaced by the Flemish L1 training set, as model Con-F1 shows, an apparent improvement was observed. By combining the Netherlandic and Flemish L1 training sets, a better WER was obtained in model Con-N1-F1. In the follow-up step, we combined

the Netherlandic L1 and L2 training set, and the WER further declined to 37.26%. Since there are some Flemish L2 data available, we then explored the combination of Flemish L1 and L2 training sets in model Con-F1-F2. Compared with model Con-N1-N2, the apparent improvement of model Con-F1-F2 may be due to the fact that the mismatch of the training set and test set is smaller in this case. Nevertheless, through the comparison between model Con-F1-F2 and Con-F2, it becomes clear that the model cannot benefit from Flemish L1 data. On the other hand, we found a WER reduction when comparing model Con-F2 with model Con-N2-F2, which indicates that the Netherlandic Dutch L2 speech data can improve the performance of the model. Through the combination of the Netherlandic L2 and the Flemish L2 training sets, model Con-N2-F2 achieved the lowest WER of 17.42%. However, even though we combined all four training sets, model Con-all performs slightly worse than model Con-N2-F2.

Table 4.4: *WERs of Con(ventional) and T(ransfer) L(earning) DNN-based ASR models for non-native Flemish Dutch learners (F2). N = Netherlandic Dutch, F = Flemish Dutch; 1 = native speech, 2 = non-native speech. The last two WERs relate to Ref(erence) models for native Dutch and Flemish.*

Model name	WER
Con-N1	77.52%
Con-F1	56.54%
Con-N1-F1	50.69%
Con-N1-N2	37.26%
Con-F1-F2	19.20%
Con-F2	19.13%
Con-N2-F2	17.42%
Con-all	18.03%
TL-N1	54.19%
TL-N1-N2	53.12%
Ref-N1	10.93%
Ref-F1	17.36%

Except for the above-mentioned conventional DNN-based models, we also explored the transfer learning-based models. Although the training sets of the two transfer learning models are the same, both are Flemish L1 speech, the training data of the pre-trained models vary. The pre-trained data of model TL-N1 is from the Netherlandic L1 dataset, while the combination of the Netherlandic L1 and L2 for the pre-trained model of model TL-N1-N2. Compared with model Con-F1, the WER of model TL-N1 is slightly lower. When we replaced the pre-trained data with the combination of the Netherlandic L1 and L2 training sets, the WER of model TL-N1-N2 decreased slightly.

4.3.1.1 10-fold cross-validation results

As mentioned above, there is an overlap in speakers (but not in utterances) if F2 data are also used for training, i.e., for models Con-F2, Con-F1-F2, Con-N2-F2, and Con-all. Therefore, we implemented 10-fold CV experiments for these models. WERs of the 10-fold experiments are illustrated in Table 4.5. The durations of the various training models in each CV experiment are approximate, around 1.61 hours, 17.6 hours, 6.4 hours, and 51.2 hours for models Con-F2, Con-F1-F2, Con-N2-F2, and Con-all, respectively. In Table 4.5 we can see that the WERs in each CV experiment vary considerably. This can be explained by the fact that the proficiency levels of the L2 learners and the recording qualities are different. Except for model Con-N2-F2, the average WERs of the other three models in Table 4.5 are close to the WERs of the corresponding models in Table 4.4. The higher average WER of model Con-N2-F2 in Table 4.5 than in Table 4.4 may be due to the more unbalanced proficiency levels of the speakers. On the whole, the average WERs in Table 4.5 are equivalent to the WERs of corresponding models in Table 4.4, which indicates that the overlap of the speakers has little effect on the performance of the proposed ASR models.

Table 4.5: *WERs of 10-fold CV experiments.*

Model name	WER	Model name	WER	Model name	WER	Model name	WER
Con-F2-1	20.57%	*Con-F1-F2-1*	23.56%	*Con-N2-F2-1*	19.53%	*Con-all-1*	21.48%
Con-F2-2	19.54%	*Con-F1-F2-2*	20.35%	*Con-N2-F2-2*	19.61%	*Con-all-2*	17.73%
Con-F2-3	15.87%	*Con-F1-F2-3*	17.46%	*Con-N2-F2-3*	15.73%	*Con-all-3*	15.80%
Con-F2-4	25.92%	*Con-F1-F2-4*	25.33%	*Con-N2-F2-4*	25.70%	*Con-all-4*	23.72%
Con-F2-5	23.75%	*Con-F1-F2-5*	25.02%	*Con-N2-F2-5*	23.19%	*Con-all-5*	21.83%
Con-F2-6	14.75%	*Con-F1-F2-6*	14.50%	*Con-N2-F2-6*	15.42%	*Con-all-6*	15.54%
Con-F2-7	16.21%	*Con-F1-F2-7*	15.93%	*Con-N2-F2-7*	17.20%	*Con-all-7*	16.70%
Con-F2-8	12.69%	*Con-F1-F2-8*	12.31%	*Con-N2-F2-8*	11.44%	*Con-all-8*	12.38%
Con-F2-9	16.03%	*Con-F1-F2-9*	15.50%	*Con-N2-F2-9*	17.09%	*Con-all-9*	14.53%
Con-F2-10	27.69%	*Con-F1-F2-10*	24.30%	*Con-N2-F2-10*	24.06%	*Con-all-10*	22.37%
Average	19.30%	*Average*	19.43%	*Average*	18.90%	*Average*	18.21%

4.3.2 GOP-based pronunciation error detection

A phone-independent threshold was applied to determine whether each phone is correct or incorrect in the present study. 50 sets of thresholds were finetuned to find the optimal threshold value. For each search, measures mentioned in section 4.2.3 were computed according to the current threshold. The step size of the thresholds

was based on the maximum and minimum GOP scores of each model. The curves in Figure 4.4 are plotted according to the results at each step.

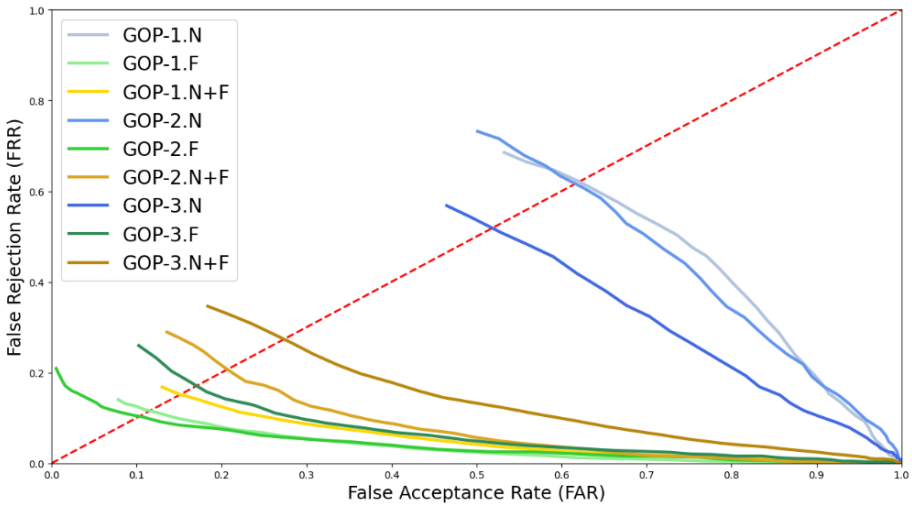

Figure 4.4: *FAR-FRR curves for the proposed GOP-based models.*

The optimal threshold for each model was obtained according to the EER. The point where the FAR and FRR cross indicates the best threshold. Table 4.6 gives the results on these thresholds. In Table 4.6, models denoted with 'N' represent the training set using Netherlandic Dutch speech data, models denoted with 'F' indicate models trained with the Flemish Dutch speech data, while 'N+F' means modeling with the combination of the Netherlandic Dutch and Flemish Dutch speech data. Considering the optimal threshold being selected at the EER point, the FAR is approximate to the FRR in each model but not the same on account of the accuracy and decimal digits of the GOP scores, as well as the results for Recall, Precision, and F1-score of CA and CR. Except for the above-mentioned measures, Cohen's kappa (κ) was also taken into account to evaluate the correlation between the training set and test set of each model. The larger the value of , the higher the correlation. Therefore, in Table 4.6, the values of of the first model in the three sets of models are the lowest. As demonstrated in Table 4.6, the second model in each set achieved the best results. The disappointing results of the first model in each set, as a result of the mismatch between the training set and test set, are plausible. What can also be clearly found in this table is the performance decline of the third model, compared to the second one, in each set. It reveals that the combination of two training sets does not benefit the model. In addition, the variability of the results in Table 4.6 is the same as in Figure 4.4.

Table 4.6: *Evaluation of the three GOP-based models for each of the three vowel pairs (see Table 4.3).*

Model name	FAR	FRR	SA	ϰ	Recall of CA	Precision of CA	F1-score of CA	Recall of CR	Precision of CR	F1-score of CR
GOP-1.N	62.13%	62.15%	37.86%	-0.2423	37.85%	37.71%	37.78%	37.87%	38.02%	37.95%
GOP-1.F	11.51%	11.55%	88.47%	0.7698	88.45%	88.43%	88.43%	88.49%	88.52%	88.51%
GOP-1.N+F	15.16%	15.22%	84.81%	0.6962	84.78%	84.75%	84.77%	84.84%	84.86%	84.86%
GOP-2.N	61.64%	61.72%	38.32%	-0.2324	38.28%	38.32%	38.36%	38.36%	38.31%	38.34%
GOP-2.F	10.25%	10.33%	89.71%	0.7946	89.67%	89.73%	89.70%	89.75%	89.69%	89.72%
GOP-2.N+F	20.69%	20.86%	79.23%	0.5843	79.14%	79.29%	79.22%	79.31%	79.17%	79.24%
GOP-3.N	51.71%	51.81%	48.24%	-0.0348	48.19%	48.14%	48.17%	48.29%	48.35%	48.32%
GOP-3.F	17.08%	17.16%	82.88%	0.6577	82.84%	82.84%	82.84%	82.92%	82.92%	82.92%
GOP-3.N+F	27.20%	27.29%	72.75%	0.4544	72.71%	72.75%	72.73%	72.80%	72.75%	72.78%

4.4 Discussion and conclusions

In this study, we have investigated several approaches to improving ASR and PED performance on learners' speech in a pluricentric language, Dutch, with speech data from two varieties, Netherlandic (dominant in the sizes of the speech resources) and Flemish (non-dominant). We cumulated speech resources to find out whether a non-dominant variety (Flemish) might benefit from speech resources from the dominant variety (Netherlandic). Our training data includes a large amount of Netherlandic Dutch native and non-native read speech data as well as a small amount of Flemish Dutch native and non-native speech data. The focus of our research questions is on modeling Flemish non-native speech data.

First, we applied recent state-of-the-art DNN-based acoustic modeling methods with different combinations of the datasets. WER is adopted to evaluate the performance of the acoustic models. In Table 4.4 we can observe three cases in which adding data from the dominant variety reduces WER: Con-F1 => TL-N1; Con-F1 => Con-N1-F1; Con-F2 => Con-N2-F2. The general conclusion for ASR is that using or adding Dutch data helps, but that putting all speech datasets together (the most complete model Con-all) does not deliver the best result. The reduction of the WERs reveals that adding N2, non-native Netherlandic, helps, but the strongest effect comes from adding F2, non-native Flemish. The results are comparable to those obtained for pathological speech in previous research (Yilmaz et al., 2016), which also obtained better performance through the combination of the Netherlandic Dutch and Flemish Dutch speech data of the patients. However, WER slightly increased in model Con-all, although the largest amount of training data was used in this case. This result may be explained by the fact that the Netherlandic Dutch data in the training set was much larger than the Flemish Dutch data, thus creating an imbalance. Furthermore, the language learners with different mother-tongue backgrounds in the Netherlandic Dutch dataset make this a more challenging task. Both factors are likely to result in a reduction in the discriminative power of the ASR model. The outcomes of the reference models indicate that the performance of the best model, Con-N2-F2, is fairly good, particularly in comparison to the Flemish Dutch references model (Ref-F1).

In the '2008 N-best challenge', native Broadcast News (BN) and Conversational Telephone Speech (CTS) were proposed for both Netherlandic Dutch and Flemish Dutch. Considering that this challenge and our work both used speech data of the CGN corpus and focused on the two Dutch standard varieties, we compared our results of the two reference ASR models with those of the challenge. According to van

Leeuwen (2013), the best WERs (obtained by Despres et al., 2009) in this challenge were 17.8% (BN_NL), 15.9% (BN_VL), 46.1% (CTS_VL), and 46.1% (CTS_VL). The WER of model Ref-N1 (10.93%) outperforms both BN_NL and CTS_NL tasks, while the WER of model Ref-F1 (17.36%) is slightly higher than that of the BN_VL task but much better than that of the CTS task in Flemish Dutch. The native speech resource of Flemish Dutch is about half of that of Netherlandic Dutch. Hence, with the same modeling procedures, it is plausible that the performance of model Ref-N1 is better than that of model Ref-F1. The durations of the training data of BN_NL, BN_VL, CTS_NL, and CTS_VL are 99.4 hours, 92.0 hours, 52.9 hours, and 64.0 hours, respectively, which is much more than the native speech data in the present study. Although the DNN-based classifiers generally show better performance, insufficient training material may lead to the WER of model Ref-F1 being higher than that of the BN_VL task. Furthermore, note that in our experiments, we used the same lexicon and language model for all tasks because we wanted to investigate the effect of changing only the acoustic models. Lexicons and language models optimized for specific tasks (i.e., for the Ref-N1 and Ref-F1 tasks) might yield lower WERs. Another study of the '2008 N-best challenge', namely Huijbregts et al. (2009), presented results that are opposite to those of the current research. Combining all the Netherlandic and Flemish Dutch data resulted in an increase in WER for their Netherlandic Dutch BN development set. Apart from the fact that conversational speech recognition is a more challenging task compared to read speech recognition, we also think that the mismatch between the BN and CTS data led to results that are contrary to those in our own study.

We furthermore trained two transfer learning-based models based on models Con-N1 and Con-N1-N2. These models might be interesting from the viewpoint of how learning a new language actually develops. With limited non-native learners' speech resources, Duan et al., (2019) used two large English and Japanese native speech corpora for multi-lingual acoustic modeling of the Japanese English learners. Lower WERs were achieved on both consonants and vowels pronunciation error diagnosis. Similar to the results of Matassoni et al. (2018), both our transfer models attained better results than the plainest conventional model, Con-F1 in our case. Moreover, model TL-N1-N2 proved to be the best by combining the Netherlandic Dutch L1 and L2 datasets, which also benefitted from the combination of different domains of data. Nevertheless, the results are disappointing when compared to the performance of Con-N1-N2. This demonstrates that even though transfer learning can improve the performance of the model to a certain extent, the model benefits more from more matching data.

Next, we explored the performance of GOP-based PED models, again combining Netherlandic Dutch and Flemish Dutch speech data. Due to the lack of the Flemish

Dutch learners' speech data with error transcriptions, we trained three sets of GOP-based models on native speech data, each set including three models using vowel pairs to simulate pronunciation errors made by Francophone learners of Dutch, with a percentage of artificial errors in each model of 50%. Several metrics were used to illustrate the detection performance of the models. For the three models in each set, we found that the Flemish (F) model outperforms the other two models, i.e., Netherlandic (N) and their combination (N+F). Even though the combination model was trained based on more data, its performance in each set is slightly worse than that of the Flemish model but much better than that of the Netherlandic model. This reveals that combining data from these two varieties of Dutch may affect performance, in this case in a negative way. The unsatisfactory result of the Netherlandic model should mainly be attributed to the mismatch between the training set and the test set. Van der Harst et al. (2014) found systematic vowel differences in the whole vowel system between the two Dutch varieties. They used discriminant analysis to distinguish 80 adult speakers of Flemish and 80 speakers of Dutch, based only on the F1 and F2 measurements of eight different monophthongal vowels. Ninety percent of the speakers were correctly classified according to variety. The patterns of variation between the varieties involved seem to be high enough to interfere with the recognition of error patterns.

The artificial errors introduced are frequent pronunciation errors made by Francophone learners of Dutch L2 (Hiligsmann & Rasier, 2006). Both /ɑ/ - /a:/ and /I/ - /i/ are real error pairs that actually occur in non-native data, while /a:/ - /i/ is a pair used as sanity check as these two vowels are maximally different. Compared to Witt (1999), the proposed model GOP-2.F achieved a better SA (89.71% against 80% with MFCC, respectively), but the remaining two models of the second set attained lower accuracies. This may also be caused by mismatches between the training and test set. Compared to the research by Kanters et al. (2009), better Recall, Precision, F1 score, and SA results were gained on our proposed 'F' model of each set (on the same test set). This probably can be explained by the fact that our proposed models are DNN-based, which makes the classification ability of the models relatively better. Among the three pronunciation errors that were introduced, the pair /a:/ - /i/ is the easiest to distinguish, and as a matter of fact, this is not a realistic mispronunciation. We see that the overall performance of the second set is the best. The models that combine the Netherlandic and Flemish datasets perform worse. This seems to reveal that distinguishing /I/ from /i/, and /ɑ/ from /a:/ is also easier when only Netherlandic Dutch data is used. The /ɑ/ - /a:/ pair performs rather close to the sanity check pair. This is confirmed by the measurements in van der Harst et al. (2014), with only a slight distinction in the /a:/ between the two varieties (F1 values). Both varieties have

a marked length distinction between the /ɑ/ and /a:/. Particularly the /I/ shows a strong distinction between the two varieties (F2 values), and this applies to the /i/ as well (again, F2 values) (Kloots et al., 2006; van de Velde et al., 2010). This variability might explain the lower performance of the /I/ - /i/ pair. These outcomes seem to demonstrate that applying PED to real non-native data is a promising avenue for future research, also in a pluricentric context.

From the results presented in this paper, we can draw the following conclusions. First, the Flemish Dutch learners' ASR model can be improved by cumulating the Netherlandic Dutch and Flemish Dutch datasets. Second, the transfer learning-based model can effectively transfer the Netherlandic Dutch knowledge to promote the Flemish Dutch learners' ASR mode, but interference problems may arise, and its performance is far from impressive. Third, although the performance of the combination of data from the two varieties of the Dutch is close to the models trained with Flemish Dutch speech data, the GOP-based PED models cannot benefit from speech data cumulation. Taken together, these results indicate that while for conventional speech recognition, Dutch-Flemish data cumulation is helpful, for more detailed analyses at the phonemic level as in the PED models combining the data reduces discriminatory power as a result of the phonological variation in the two varieties of Dutch.

Finally, we should consider applying our pluricentric approach with articulatory features. Compared to the conventional acoustic features like MFCC, articulatory features are more closely related to speech production and have the advantage of being less affected by environmental noise.


Hello
你好

Chapter 5

Articulatory-enhanced End-to-End Mispronunciation Detection and Diagnosis

This chapter is based on the following publications:

Wei, X., Dong, W., Cucchiarini, C., van Hout, R., & Strik, H. (2025). Leveraging Articulatory Information to Enhance End-to-End Models for L2 Mispronunciation Detection and Diagnosis. *10th Workshop on Speech and Language Technology in Education (SLaTE)*, 76–80. https://doi.org/10.21437/slate.2025-16

Wei, X., Cucchiarini, C., van Hout, R., & Strik, H. (2025). Articulatory-Enhanced Mispronunciation Detection and Diagnosis Models: A Multi-dimensional Error Analysis. *10th Workshop on Speech and Language Technology in Education (SLaTE)*, 31–35. https://doi.org/10.21437/slate.2025-7

5.1 Leveraging Articulatory Information to Enhance End-to-End Models for L2 Mispronunciation Detection and Diagnosis

Pronunciation variability in second language (L2) speech poses crucial challenges for mispronunciation detection and diagnosis (MDD) models. This study leverages articulatory features (AFs) to enhance End-to-End (E2E) MDD models and provide better diagnostic feedback. We investigate AF integration under two distinct output representation frameworks: phoneme-based (PHN) and articulatory-based (ART), using both customized Conformer-based models and fine-tuned XLSR models. The experimental results indicate that AF integration consistently improves detection accuracy and reduces diagnostic error rates across models and frameworks. Notably, combining AFs with XLSR embeddings within the ART framework achieves the lowest diagnostic error rates, enabling the most informative subsegmental feedback. These findings highlight the value of articulatory-informed pre-trained models in developing more accurate and feedback-rich L2 pronunciation training tools.

5.1.1 Introduction

The rising demand for second language (L2) learning has drawn greater focus in computer-assisted language learning (CALL) research. Mispronunciation detection and diagnosis (MDD), a key aspect of CALL, helps identify pronunciation errors and offer feedback to L2 learners for more effective improvement. MDD models can operate at multiple levels: supra-segmental (e.g., pitch accent (Ren et al., 2004), lexical stress (Höhle & Weissenborn, 2003), etc.), segmental (e.g., substitution of phonemes), and subsegmental (e.g., voicing feature activated for a canonical unvoiced phone (Stevens, 2000)) levels.

In recent decades, significant research efforts have been dedicated to enhancing phoneme-based MDD models, which continue to be the most prevalent approach at the segmental level. Broadly, these methodologies can be classified into two main categories. The first category comprises scoring models, which correlate a reference transcript with its corresponding pronunciation and are typically followed by the application of heuristic scoring functions, such as Goodness of Pronunciation (Witt & Young, 2000). For these models, establishing a threshold is essential in determining the correctness of the pronunciation (Hu et al., 2015b). The second category consists of classification-based approaches, which first convert input speech into recognized phonemes via classifiers and then align these phonemes with canonical phonemes to deliver feedback. This can be accomplished through techniques such as conventional

neural network-based approaches (Li et al., 2016a; Nazir et al., 2019) and End-to-End (E2E) based models (Gao et al., 2024; Ye et al., 2022).

A major limitation of previous methods is their reliance on pronunciation errors observed during training or predefined by manual rules. However, many L2 mispronunciations involve subtle distortion errors that slightly diverge from canonical sounds (Yoon et al., 2010). For instance, Japanese learners of English tend to produce a sound between aspirated /p/ and unaspirated /b/, rather than consistently replacing /p/ with /b/. Additionally, certain phonemes in the target language may be absent in the learners' native language, as seen in the confusion between 'lead' and 'read' due to the lack of /r/ in Japanese (Riney & Anderson-Hsieh, 1993). These phenomena present significant difficulties for phoneme-based MDD models, as they would require acoustic models to cover a broad set of phonemes, including both L1 sounds and their variants, which is inherently impractical.

Moreover, segmental-level feedback has shown limited effectiveness in improving L2 pronunciation. Studies suggest learners benefit more from direct, subsegmental instructions addressing articulatory deviations (Bälter et al., 2005). For instance, Japanese learners may refine their /r/ production more effectively when instructed to retract their tongue from an /l/-like articulation. Motivated by such advantages, recent work has incorporated subsegmental information into L2 learning (Chen et al., 2022; Mao et al., 2018b; Siriwardena et al., 2022; Sun et al., 2012; Zhong et al., 2024), including the use of articulatory features (AFs), which are robust under noisy conditions and can significantly reduce word error rate (Kirchhoff et al., 2002). To avoid the difficulty of collecting physical articulatory data, rule-based acoustic-to-articulatory mappings have been employed to link phonemes to corresponding articulators (Li et al., 2016b).

In addition to the modeling efforts discussed, a persistent challenge in MDD lies in the limited availability of annotated L2 data. Although pre-trained models such as Wav2Vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023) alleviate data scarcity, existing frameworks that rely solely on acoustic or phoneme-level information still struggle to represent the full spectrum of phonetic variability and nuanced articulation patterns found in L2 speech. To address these limitations, we propose two novel methods for incorporating AFs into E2E MDD modeling: one applied to customized Conformer-based models and the other to the fine-tuned Wav2Vec 2.0 models. As an additional key novel aspect, we implement both methods within two output frameworks: a phoneme-based (PHN) framework and an articulatory-based (ART) framework. The main innovation aspects of our study

lie in the systematic incorporation of AFs in both methods, aiming to improve performance across diverse configurations and to assess whether subsegmental representations yield more interpretable, linguistically grounded feedback. To this end, we address two main research questions: RQ5.1.1) To what extent do AFs enhance the performance of the proposed MDD models? RQ5.1.2) At which level, segmental or subsegmental, is the integration of AFs more advantageous?

5.1.2 Method

5.1.2.1 Speech material

This study uses two speech corpora: Librispeech (Panayotov et al., 2015) and L2-ARCTIC (Zhao et al., 2018). Librispeech is an English native corpus containing approximately 1000 hours of read speech sourced from audiobooks available through the LibriVox project (Kearns, 2014). A 100-hour subset of clean speech was selected to train the AF classifiers. The L2-ARCTIC corpus features recordings from 24 speakers with diverse linguistic backgrounds (e.g., Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese), serving as the English L2 data. From the total 27 hours of speech, 3.5 hours of manually annotated data were used for MDD modeling. The L2 data were randomly split into three subsets: 12 speakers for training, 6 for validation, and the remaining 6 for testing, ensuring no overlap of speakers or utterances among the subsets. Note, unlike the phoneme labels in the other chapters, which are presented in IPA format, the phoneme labels in this chapter are represented in SAMPA. This choice was made because many previous studies using the L2-ARCTIC corpus adopted SAMPA-based phoneme labels, and using the same format facilitates direct comparison with those studies.

5.1.2.2 Extraction of AFs

The overall architecture of this study is illustrated in Figure 5.1.1. The initial module comprises six DNN-based classifiers built in Kaldi (Povey et al., 2011), which utilize 39-dimensional MFCC+Δ+ΔΔ vectors as input features, following (Wei et al., 2017). A comparison of input features for the DNNs, specifically contrasting FBank pitch with MFCC, demonstrates that MFCC-based models consistently outperform FBank pitch models, leading to their selection for modeling. Following forced alignment procedures, context-dependent articulatory labels were applied separately to train the corresponding DNNs, each comprising six hidden layers with 2048 sigmoid units per layer. The output layer utilizes the softmax function to generate frame-level posterior probabilities. These individual classifier outputs are subsequently concatenated frame by frame through the append module to construct the final AF representations.

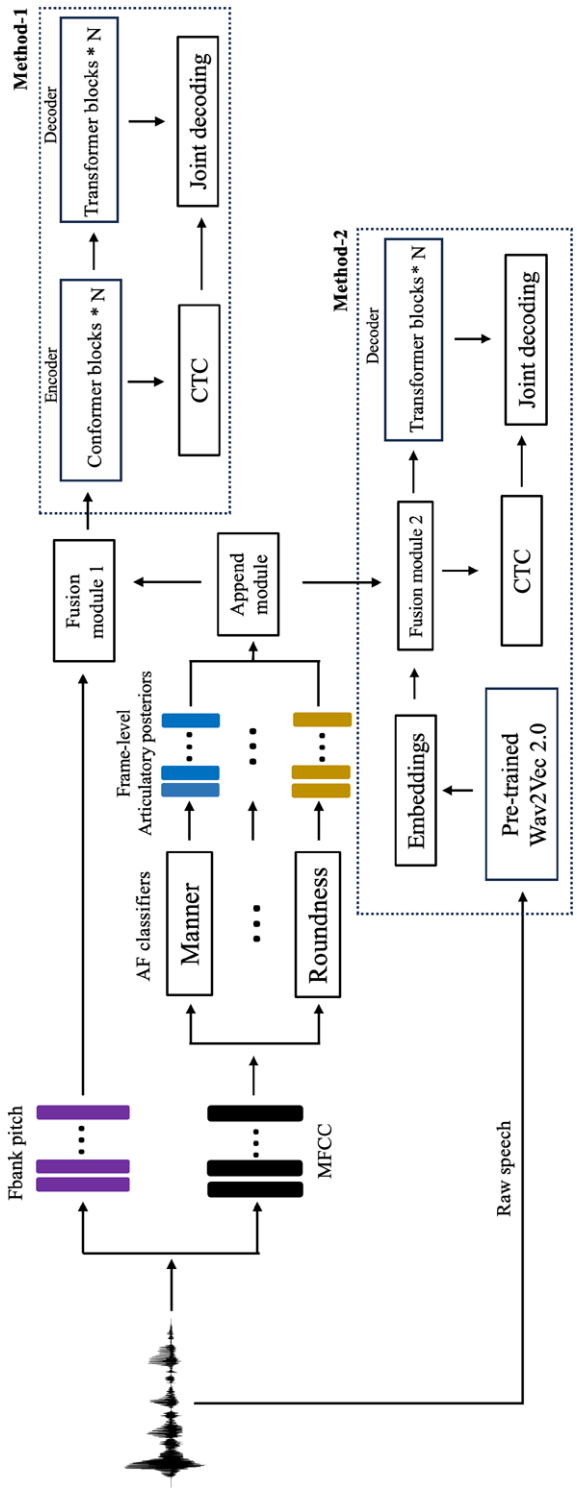

Figure 5.1.1: *Illustration of the two proposed methods.*

Analysis of pronunciation errors in the L2 datasets revealed systematic patterns in both vowel and consonant productions, prompting the development of three articulatory categories (ACs) for each phoneme type, adapted from the definition in Frankel et al. (2007). For consonants, the categories are 'Manner', 'Place', and 'Voicing', the categories for vowels are 'Backness', 'Height', and 'Roundness'. A detailed description of these ACs and corresponding attributes is provided in Table 5.1.1.

Table 5.1.1: *List of articulatory categories (ACs) and their associated features (AFs).*

Articulatory category	Articulatory feature
Manner	Affricate, Fricative, Nasal, Stop, Approximant
Place	Alveolar, Bilabial, Dental, Glottal, Labiodental, Palatal, Post-alveolar, Velar
Voicing	Voiced, Unvoiced
Backness	Front, Central, Back, Back2front
Height	High, Middle, Low, Low2high
Roundness	Rounded, Unrounded, Rounded2unrounded

5.1.2.3 Proposed frameworks and models

This study examines two modeling methods employed within distinct frameworks: a phoneme-based (PHN) framework at the segmental level and an articulatory-based (ART) framework at the subsegmental level. In the PHN framework, each acoustic unit corresponds to a unique symbol. In contrast, the ART framework employs articulatory labels that can map to multiple distinct phonemes, and a single phoneme may be associated with various labels across different articulatory categories. For instance, the phoneme /AH/ could be connected to articulatory labels like 'central', 'middle', and 'unrounded', as specified in Table 5.1.1. Within each framework, five models were evaluated, categorized into two primary methods (Method-1 and Method-2) based on their input features and training strategy.

The first method (Method-1) uses custom-designed E2E models based on a hybrid CTC-Attention architecture, which jointly leverage CTC and attention mechanisms for training and decoding, eliminating the need for forced alignment. These models consist of three key components: a Conformer encoder, composed of 12 Conformer blocks with 1024-dim linear units, processes input features, which are initially down-sampled through a two-dimensional series of Conformer encoder layers for feature extraction. Compared to Transformer, Conformer is more effective at modeling local feature patterns (Gulati et al., 2020), which is particularly beneficial for mispronunciation

detection tasks. The second component is a Transformer-based decoder, which includes 4 Transformer blocks with 1024-dim linear units and jointly processes encoder outputs and speech embeddings of the text, followed by a CTC module for auxiliary alignment. Under this method, three models were built for each framework: a baseline using raw speech signals (RS); a baseline employing 83-dimensional FBank pitch features (FP); and a proposed model (M1), which incorporated FBank pitch with AFs. Since AFs and FBank pitch share the same temporal resolution, fusion module 1 performs a direct frame-by-frame concatenation of the two features. Specifically, given two input matrices of shape $T \times D_1$ and $T \times D_2$ representing AFs and FBank pitch features respectively (where T is the number of frames, and D_1,D_2 are their respective feature dimensions), the fusion results in a single $T \times (D_1+D_2)$ matrix.

The second method (Method-2) leverages the multilingual pre-trained model XLSR (Babu et al., 2021) to handle the linguistic diversity of L2 speakers. Both baseline and proposed models process raw speech inputs. The baseline (FT) directly fine-tunes XLSR (https://huggingface.co/facebook/wav2vec2-large-xlsr-53) with the training process optimized using the validation set to prevent overfitting. In contrast, our proposed model (M2) integrates speech embeddings extracted from the encoder with AFs before feeding them into the Transformer-based decoder and CTC module. Fusion module 2 addressed temporal mismatches between AFs and speech embeddings by first zero-padding the shorter sequence along the time axis to match the longer one. Given AFs and embeddings of shapes $T_1 \times D_1$ and $T_2 \times D_2$ ($T_1 \neq T_2$), padding align both to $max(T_1, T_2)$ frames. The two matrices are then concatenated along the feature dimension to yield a $max(T_1, T_2) \times (D_1 + D_2)$ representation for frame-wise fusion.

5.1.3 Results

5.1.3.1 Consistency results of ACs and AFs

For each AF frame, the label with the highest posterior probability was selected and compared against the reference to compute frame-level consistency rates, as illustrated in Table 5.1.2. Results indicate that the consistency rate across the three L2 datasets ranged from 74% to 82%, suggesting that most AFs were accurately extracted. While rates vary across ACs, they remain consistent within each AC across the three datasets.

To evaluate the consistency performance of each AF, we carried out an additional analysis to examine their consistency rates, as illustrated in Figure 5.1.2. Certain features, such as 'Front' and 'Unvoiced', demonstrate consistency rates exceeding 80%, whereas other features, such as 'Rounded2unrounded', show significantly lower consistency rates. In alignment with the findings in Table 5.1.2, the consistency rates of both articulatory categories and features appear to be affected by sample size. For instance, ACs like 'Roundness' and AFs such as 'Fricative' achieve higher consistency rates. Conversely, ACs like 'Place' show lower consistency rates mightily due to fewer samples per AF, while AFs like 'Back2front', linked to fewer phonemes, also suffer from limited training data. As reported in Zhao et al. (2018), the phoneme /OY/ accounts for only 0.3% of the data, which contributes to the diminished consistency rates for AFs such as 'Back2front' and 'Rounded2unrounded'. These results suggest that, while there is room for improvement, the process of AF extraction is reasonably reliable.

Table 5.1.2: *Consistency rates of ACs in the L2 datasets.*

	Training (%)	Validation (%)	Testing (%)
Manner	76.35	77.92	76.41
Place	74.06	75.58	74.66
Voicing	78.13	78.90	78.50
Backness	76.71	78.72	76.99
Height	75.21	76.41	75.55
Roundness	80.32	81.75	80.28

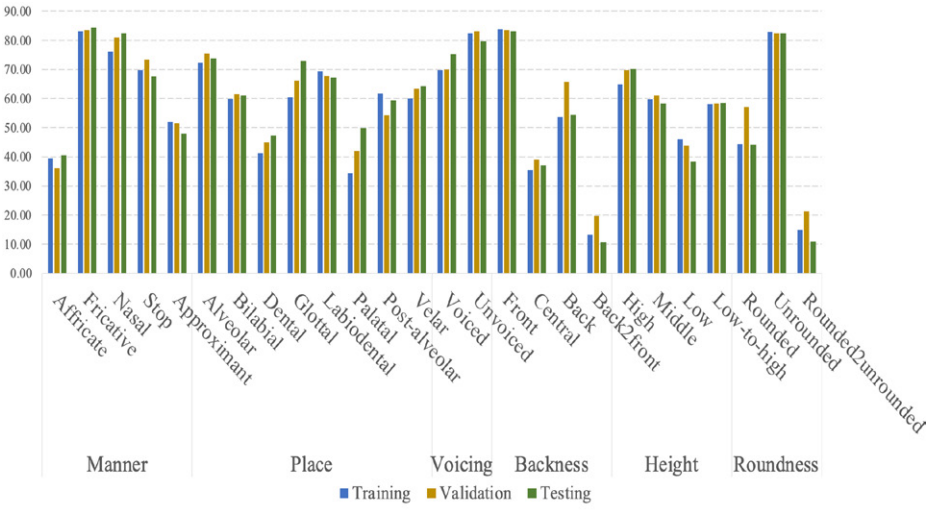

Figure 5.1.2: *Consistency rate of AFs in the L2 datasets.*

5.1.3.2 Evaluation of MDD models

This section reports the evaluation results of the proposed and baseline MDD models on the L2-ARCTIC test set, as shown in Table 5.1.3. Each model produces either a phoneme or an articulatory label sequence, which is then aligned with human annotations for evaluation. Performance is measured using phoneme error rate (PER) for recognition quality, detection accuracy (DA), false rejection rate (FRR), false acceptance rate (FAR), and diagnostic error rate (DER) for mispronunciation detection. We also report precision, recall, and F1-score on correctly rejected labels.

5.1.3.2.1 Comparison of proposed and baseline models

We first evaluated the baseline models and compared them with the proposed models that incorporate AFs. In Method-1, models trained on raw speech (PHN-RS and ART-RS) generally out-performed those trained on FBank pitch (PHN-FP and ART-FP) across both frameworks, although specific metrics, such as Recall in PHN-RS, were slightly lower than in PHN-FP. The proposed models that integrated AFs with FBank pitch (PHN-M1 and ART-M1) outperformed their respective baselines, as reflected in the performance gains of ART-M1 over ART-RS and ART-FP. For Method-2, the baseline models (PHN-FT and ART-FT), fine-tuned from a pre-trained XLSR model, achieved better performance than the Method-1 models, likely due to the large-scale training data used for pretraining. The proposed models in Method-2 (PHN-M2 and ART-M2) further improved upon their baselines, indicating the effectiveness of incorporating AFs in the fine-tuning process.

5.1.3.2.2 Comparison of MDD frameworks: PHN vs. ART

When comparing the two MDD frameworks under the proposed models, the ART framework exhibited notable performance differences relative to the PHN framework. In method-1, although model ART-M1 showed a slight decrease in DA and FRR compared to PHN-M1, its DER was significantly reduced, dropping by 27.98% from 41.46% (PHN-M1) to 29.86% (ART-M1). Moreover, the DER achieved by ART-M1 (29.86%) is substantially lower than reported in a previous customized E2E phoneme-based MDD model (36.33%; Dong et al., 2023). In Method-2, the comparison between PHN-M1 and ART-M2 yielded more balanced results: PHN-M2 outperformed ART in four metrics, while ART-M2 surpassed PHN-M2 in the remaining four. Importantly, DER improved markedly in ART-M2, decreasing relatively by 25.88% from 25.12% (PHN-M2) to 18.62%.

Table 5.1.3: *Results of MDD models in phoneme-based (PHN) and articulatory-based (ART) frameworks. RS: raw speech signals; FP: FBank pitch; M1: proposed model in Method-1; FT: finetune XLSR; M2: proposed model in Method-2.*

			PER	DA	DER	FAR	FRR	Precision	Recall	F1
Phoneme	Method-1	PHN-RS	22.14	81.74	44.63	57.84	9.92	47.22	42.16	44.55
		PHN-FP	23.11	81.02	46.40	57.76	10.81	45.15	42.24	43.65
		PHN-M1	21.77	82.52	41.46	51.59	10.30	49.76	48.41	49.08
	Method-2	PHN-FT	17.57	86.28	33.99	**44.74**	7.19	61.82	**55.26**	58.36
		PHN-M2	17.12	**86.88**	25.12	45.37	6.33	64.51	54.63	**59.16**
Articulatory	Method-1	ART-RS	18.88	82.40	28.32	56.91	9.32	49.33	43.09	46.00
		ART-FP	20.16	81.83	33.22	55.37	10.34	47.63	44.63	46.08
		ART-M1	19.85	82.48	29.86	50.09	10.67	49.64	49.91	49.78
	Method-2	ART-FT	14.69	85.98	21.42	52.97	5.82	62.98	47.03	53.85
		ART-M2	**13.61**	86.75	**18.62**	54.17	**4.63**	**67.57**	45.83	54.62

5.1.3.2.3 Error analysis: vowels vs. consonants

To explore possible reasons for the minor decline observed in certain ART model metrics, we analyzed the FAR and FRR for the five frequent vowel and consonant substitution errors in the L2-ARCTIC for the proposed models in Method-1 (PHN-M1 and ART-M1), as shown in Table 5.1.4, focusing exclusively on substitution errors, as they far outnumber insertion and deletion errors, which were excluded from the calculation of the metrics reported in Table 5.1.4. Results indicate that AFs are more effective in detecting vowel errors than consonant errors, similar to findings in (Shahin & Ahmed, 2024). This imbalance may affect the overall performance since consonants account for 60.8% of the phoneme distribution in the L2-ARCTIC corpus. Another possible explanation could be a misalignment between the input and output representations of the model. While PHN models predict phoneme-level targets that align with the granularity of AFs, ART models yield articulatory labels of a relatively coarser nature, potentially diminishing the models' capacity to identify detailed mispronunciation patterns. Consequently, this mismatch in representation might restrict the discriminative efficacy of the ART framework in certain situations. In addition, as shown in Table 5.1.4, some mispronunciation pairs, such as "Z/S" and "DH/D", yielded noticeably better results than others. This may be partly attributed to the distribution of sample counts and possibly to differences in pronunciation quality, which may have indirectly influenced model performance.

Table 5.1.4: *FAR and FRR of the most common vowel and consonant substitution errors. Values in bold denote lower FAR/FRR in each pair of vowels and consonants.*

Vowels					**Consonants**				
	Phoneme		**Articulatory**			**Phoneme**		**Articulatory**	
	FAR	FRR	FAR	FRR		FAR	FRR	FAR	FRR
IH/IY	64.93	12.51	**58.96**	**5.72**	Z/S	28.38	**34.00**	**26.35**	36.35
OW/AO	76.00	7.49	**58.40**	**3.75**	DH/D	**27.87**	58.17	28.45	**57.52**
EY/EH	77.86	6.90	**54.20**	**5.95**	P/B	**68.85**	10.42	75.41	**9.98**
AH/AO	74.32	**11.48**	**67.57**	11.77	V/F	**48.28**	**12.27**	62.07	15.51
AE/AA	**60.61**	8.04	66.67	**7.02**	N/NG	88.89	4.48	**77.78**	**3.39**

5.1.4 Discussion and conclusions

This study introduced methods to integrate AFs with acoustic features and speech embeddings, enhancing E2E MDD models. The developed methods specifically target the pronunciation variability typical of English L2 learners and were implemented within both phoneme-based (segmental) and articulatory-based (subsegmental) frameworks.

Results presented in Table 5.1.3 reveal consistent performance improvements through AF integration across methods and frameworks. These findings address RQ5.1.1 and hold particular significance for L2 learning applications, where enhanced DA improves error identification while reduced DER yields more precise corrective feedback. Most notably, the ART framework exhibits superior performance by operating at the subsegmental level, enabling more effective capture of subtle distortion errors typically overlooked by segmental models.

To address RQ5.1.2, our comparative analysis of the segmental (PHN) and subsegmental (ART) frameworks indicates that AF integration is more advantageous at the subsegmental level. As specified in Section 5.1.3.2.2, the ART framework is inherently better suited for capturing subtle errors and aligns more closely with the need for detailed guidance for L2 learners. Specifically, within Method-2, which utilizes embeddings from a fine-tuned XLSR model, AF-enhanced models (PHN-M2 and ART-M2) outperform their non-AF counterparts, confirming the benefits of combining AFs with pre-trained representations. Although PHN-M2 and ART-M2 perform comparably on some metrics, ART-M2 delivers a substantial 25.88% relative reduction in DER (25.12% vs. 18.62%). This highlights the strength of AFs within the ART framework when integrated with XLSR embeddings, especially for addressing distortion errors through subsegmental feedback correlated with articulatory movements.

In conclusion, this study demonstrates that integrating AFs markedly enhances E2E-based MDD models, particularly when utilized with powerful pre-trained representations such as fine-tuned XLSR, as evidenced by ART-M2. The ART framework, especially ART-M2, shows clear superiority in DER reduction. This underscores the prospective capability of combining AFs, pre-trained models, and the ART framework to yield enhanced outcomes and more insightful, subsegmental feedback for L2 learners, though additional optimization and integration could enhance real-world applications.

While AF extraction is generally reliable, inconsistencies in rare ACs indicate room for improvement and may hinder the detection of subtle errors. The coarse granularity of articulatory labels further contributes to this limitation. Future work may enhance AF consistency through larger datasets or pretrained models, and address granularity mismatch by exploring hybrid outputs that combine phoneme and AF-level targets. AFs could also be used independently or integrated with other models (Dong et al., 2023), broadening their applicability for diverse L2 learners. In addition, future research could incorporate evaluation against expert human ratings to assess the extent to which model predictions align with human judgments, thereby providing further validation of the diagnostic effectiveness of the models.

5.2 Articulatory-enhanced Mispronunciation Detection and Diagnosis Models: A Multi-dimensional Error Analysis

This study presents a multi-dimensional error analysis of integrating articulatory features (AFs) into End-to-End (E2E) models for mispronunciation detection and diagnosis (MDD). We examine two output representation frameworks: phoneme-based (PHN) and articulatory-based (ART), employing both customized Conformer-based models and fine-tuned Wav2Vec 2.0 models. Experimental results reveal that AF-integrated ART models demonstrate enhanced capability in identifying common mispronunciations, thereby lowering diagnosis error rates. In contrast, PHN models maintain a slight advantage in overall detection accuracy but provide less articulatory insight. Additional analysis reveals key challenges, such as degraded detection accuracy on medium-length utterances and inter-speaker variability. These findings emphasize critical trade-offs between detection and diagnosis, while proposing practical considerations for building more accurate and learner-aware L2 pronunciation assessment systems.

5.2.1 Introduction

Mispronunciation detection and diagnosis (MDD) plays a pivotal role in computer-assisted language learning (CALL), providing second language (L2) learners not only accurate detection outcomes but also precise feedback to improve their pronunciation proficiency (Chen & Li, 2016; Cucchiarini & Strik, 2017; Neri et al., 2008). Conventional MDD systems generally rely on complex, multi-stage pipelines, such as forced alignment and pre-defined acoustic models (Hu et al., 2015a; Nazir et al., 2019). Recent advances in deep learning have shifted the focus toward end-to-end (E2E) models, which directly learn mappings from acoustic inputs to pronunciation labels, simplifying system framework and improving performance (Wang et al., 2022a; Wu et al., 2021; Yan et al., 2016).

A growing line of research investigates how to further enhance these models, particularly in challenging L2 learning contexts marked by high phonetic variability and learner-specific deviations (Tu et al., 2018). One promising avenue is the integration of articulatory features (AFs), which represent speech based on the physical configurations of the vocal tract (e.g., voicing, place, and manner of articulation) (Chen et al., 2022; Mao et al., 2018b; Shahin & Ahmed, 2004). Unlike raw acoustic features, AFs are more interpretable and linguistically grounded, offering robustness against speaker variability, a particularly valuable attribute in L2 scenarios characterized by divergent first language (L1) backgrounds and varying proficiency levels (EI Kheir et al., 2024;

Khanal et al., 2021). Moreover, AFs have the potential to support fine-grained feedback, such as subsegmental articulatory guidance (e.g., lip rounding or tongue position), which is essential for effective pronunciation training (Li et al., 2016b; Yan et al., 2023).

Prior studies have demonstrated the general benefits of AFs in enhancing MDD accuracy, yet a systematic understanding of their impact across different E2E frameworks remains limited. In particular, the comparative effectiveness of phoneme-based (PHN) models versus articulatory-based (ART) models, which operate at distinct representational levels, has yet to be thoroughly examined. Of greater concern, despite notable gains in overall accuracy, critical factors such as utterance length (Cumbal et al., 2021; Lin et al., 2023) or inter-speaker variability (Kheir et al., 2024) remain underexplored, even though they may significantly impact model behavior. These nuanced patterns are often hidden by coarse-grained evaluation methods, underscoring the need for more detailed, fine-grained analysis.

This study addresses these critical gaps by conducting a comprehensive and multi-dimensional error analysis of E2E MDD models enhanced with AFs. Departing from prior works that predominantly target accuracy improvements or proposing novel architectures, we place our emphasis on understanding model behavior across a broad design space, including different modeling paradigms (customized Conformer-based E2E model vs. fine-tuned Wav2Vec 2.0), feature configurations (with vs. without AFs), and output representations (PHN vs. ART). The key novelty of this study lies in its systematic and comparative error analysis spanning multiple architecture choices, feature sets, and representation levels, an area still underexplored in existing MDD research. To this end, we carry out a structured investigation into how factors such as utterance length, speaker variability, and mispronunciation types impact both detection performance and diagnostic precision. Through these analyses, we seek to uncover latent patterns and model limitations that are routinely overlooked, thereby offering practical insights for designing more robust and learner-aware MDD systems. Based on this objective, we pose the following overarching research question: Can in-depth error analysis of AF-enhanced E2E MDD models yield meaningful insights into their performance and behavior patterns? To explore this question, we investigate two sub-questions:

RQ5.2.1): Which performance bottlenecks and behavioral trends emerge from in-depth error analysis of AF-enhanced E2E MDD models?

RQ5.2.2): How do different output representations (PHN vs. ART) influence the relationship between detection accuracy and diagnostic precision?

5.2.2 Experimental setup for error analysis

5.2.2.1 Speech material

Two speech corpora were employed in this study: Librispeech (Panayotov et al., 2015) and L2-ARCTIC (Zhao et al., 2018). Librispeech involves approximately 1,000 hours of read English speech, originally sourced from audiobooks publicly released via the LibriVox project (Kearns, 2014). A 100-hour subset of clean speech was selected to train the AF classifiers, enabling robust acoustic-to-articulatory mapping. The L2-ARCTIC corpus contains read-aloud English speech produced by L2 speakers with diverse language backgrounds. The 3.5 hours of manually annotated data out of 27 hours of speech were selected to train the MDD models. The selected L2 data was randomly partitioned into training (12 speakers), validation (6 speakers), and test sets (6 speakers), with speaker and utterance independence across all splits. The six speakers in the test set include three native Vietnamese speakers (PNV, THV, and TLV), two native Hindi speakers (RRBI and SVBI), and one native Arabic speaker (SKA). Two are male (RRBI and TLV), and the other four are female.

5.2.2.2 Models

Following the concept outlined in (Frankel et al., 2007; Morshed & Hasegawa-Johnson, 2022), three AF categories were defined for vowels: Backness (e.g., Front, Central, Back, Back2front), Height (e.g., High, Middle, Low, Low2high), and Roundness (e.g., Rounded, Unrounded, Rounded2unrounded), and three for consonants: Manner (e.g., Affricate, Fricative, Nasal, Stop, Approximant), Place (e.g., Alveolar, Bilabial, Dental, Glottal, Labiodental, Palatal, Post-Alveolar, Velar), and Voicing (e.g., Voiced/Unvoiced). An independent DNN-HMM classifier was trained on 39-dimensional MFCC features extracted from the selected subset of Librispeech for each of the six categories. Each classifier consisted of six hidden layers, each consisting of 2048 sigmoid units, and produced frame-level posteriors through a softmax layer. These posteriors represent probability distributions over the possible values within each respective AF category (e.g., the Voicing classifier outputs posterior probabilities for *Voiced* and *Unvoiced*). Finally, the posteriors from all six classifiers were concatenated to construct a composite AF vector for each frame.

This study investigates two distinct E2E modeling methods for MDD. The first method constructs three custom E2E models that share a unified architecture, comprising a Conformer-based encoder, a Transformer-based decoder, and a Connectionist Temporal Classification (CTC) module. The key difference among these models lies in the input features: raw speech (RS) and 83-dimensional FBank pitch (FP) features serve as baselines, while the proposed M1 model incorporates a

frame-by-frame fusion of FP and AFs. To be specific, given two input feature matrices of shape $T \times D_1$ and $T \times D_2$ (where T is the number of frames of the features, and D_1,D_2 are their respective feature dimensions), fusion yield a matrix of shape $T \times (D_1+D_2)$. The second method involves fine-tuning XLSR (Babu et al., 2021), a multilingual extension of the pre-trained Wav2Vec 2.0 model (https://huggingface.co/facebook/wav2vec2-large-xlsr-53). In this setup, the baseline model (FT) utilizes only speech embeddings from the encoder, whereas the proposed M2 model integrates AFs with these embeddings before feeding them into a Transformer decoder and CTC for joint inference. These two methods yield five core model configurations: RS, FP, M1, FT, and M2. While the primary objective is segment-level prediction (i.e., phoneme transcription), this study additionally investigated subsegmental modeling by mapping phonemes to their corresponding articulatory labels, as defined by the AF categories. Accordingly, each model configuration is evaluated under two output representation frameworks: the phoneme-based (PHN) framework, which directly predicts phonemes, and the articulatory-based (ART) framework, which leverages articulatory labels for subsegmental analysis. In total, ten models are evaluated, covering all five configurations across both PHN and ART frameworks.

5.2.2.3 Evaluation metrics

To perform a multi-dimensional error analysis, a set of metrics targeting both detection and diagnostic precision is considered. For correctly pronounced speech, predictions were categorized as either Correct Acceptance (CA) or False Rejection (FR). For mispronounced speech, outcomes were classified as Correct Rejection (CR) or False Acceptance (FA). CR cases were further divided based on diagnostic precision: Correct Diagnosis (CD) refers to accurately identifying the specific mispronunciation, while Diagnosis Error (DE) denotes a misidentified error. Based on these classifications, the following metrics were computed: Detection Accuracy (DA), False Acceptance Rate (FAR), False Rejection Rate (FRR), Diagnosis Error Rate (DER), and Matthews Correlation Coefficient (MCC), as defined below:

$$DA = (CA + CR)/(CA + CR + FA + FR) \tag{5.2.1}$$

$$FAR = FA/(CR + FA) \tag{5.2.2}$$

$$FRR = FR/(CA + FR) \tag{5.2.3}$$

$$DER = DE/(CD + DE) \tag{5.2.4}$$

$$MCC = \frac{CA * CR - FA * FR}{\sqrt{(CA + FR) * (CA + FA) * (CR + FA) * (CR + FR)}} \tag{5.2.5}$$

5.2.3 Results

5.2.3.1 Overall model performance evaluation

Table 5.2.1 demonstrates the mean and standard deviation (SD) of all models concerning key evaluation metrics, calculated by averaging results across the six speakers. AF-enhanced models (M1 and M2) generally outperform their respective feature-based baselines (FP and FT) across PHN and ART frameworks, supporting the effectiveness of AFs in enhancing DA and MCC.

Table 5.2.1: *Mean (SD) performance across speakers for all evaluated models.*

Model	DA	MCC	FAR	FRR
PHN-RS	81.75(1.33)	0.33(0.08)	57.65(6.60)	9.76(2.71)
PHN-FP	81.03(1.36)	0.32(0.08)	57.67(6.02)	10.62(3.15)
PHN-M1	82.52(1.41)	0.38(0.09)	52.03(5.42)	10.14(2.84)
PHN-FT	86.29(1.22)	0.47(0.10)	47.31(8.89)	7.14(1.25)
PHN-M2	86.90(1.24)	0.48(0.11)	48.23(9.42)	6.30(0.81)
ART-RS	82.41(1.59)	0.35(0.10)	57.53(6.10)	9.16(2.87)
ART-FP	81.83(1.23)	0.34(0.09)	55.93(6.27)	10.16(3.06)
ART-M1	82.48(1.39)	0.38(0.10)	50.98(6.84)	10.49(3.07)
ART-FT	85.99(1.87)	0.44(0.10)	54.74(8.48)	5.76(1.15)
ART-M2	86.76(2.30)	0.46(0.09)	55.85(7.37)	4.60(0.77)

Table 5.2.2 presents the paired t-test results, including p-values and Cohen's d (effect size), computed across six speakers per model. The results suggest that the improvements from AF integration were often statistically significant and practically meaningful for various metrics in both frameworks.

Table 5.2.2: *Statistical significance (p) and effect size (d) for DA and MCC across key model comparisons.*

Comparison	DA	MCC
	p/d	*p/d*
PHN-M1 vs. PHN-FP	0.0045/1.9985	0.0026/2.2637
PHN-M2 vs. PHN-FT	0.0183/1.4062	0.1412/0.7128
ART-M1 vs. ART-FP	0.1158/0.7759	0.0059/1.8708
ART-M2 vs. ART-FT	0.0188/1.3975	0.0180/1.4142

5.2.3.2 Speaker-specific analysis

Figure 5.2.1 illustrates the DA heatmaps for six speakers (the six rows) across all models. Across both frameworks (PHN and ART), all speakers generally achieve higher DA scores in the proposed models M1 and M2 (compare columns 3 to 2, 5 to 4, 8 to 7, and 10 to 9). Despite these general trends, however, noticeable inter-speaker variability in DA remains evident. For instance, although ART-M1 generally outperforms ART-FP, speaker RRBI shows a marginally higher DA in ART-FP (82.74%) compared to ART-M1 (82.64%).

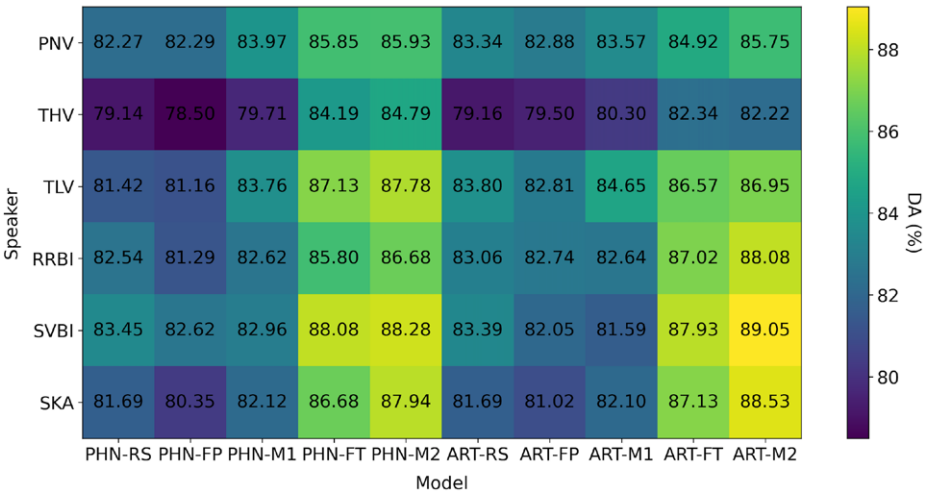

Figure 5.2.1: *Speaker-wise detection accuracy (%) across all evaluated models. (Darker indicates lower DA).*

Table 5.2.3: *Distribution (number/percentage) of mispronunciation by speakers.*

Speaker	Sub.	Ins.	Del.	Total
PNV	622/12.32	16/0.32	246/4.87	884/17.51
THV	903/17.72	67/1.31	406/7.97	1376/27.00
TLV	776/15.28	43/0.85	437/8.60	1256/24.73
RRBI	451/9.03	47/0.94	65/1.30	563/11.27
SVBI	426/8.61	28/0.57	89/1.80	543/10.98
SKA	464/9.14	116/2.29	59/1.16	639/12.59

The speaker-wise DA heatmap, along with the relatively higher mispronunciation rates of speakers THV (27.00%) and TLV (24.73%), as reported in Table 5.2.3, demonstrates a notable discrepancy between speakers. Specifically, THV attained the lowest DA scores across all models, whereas TLV achieved substantially higher

values. To further examine this unexpected divergence, we first computed FAR, FRR, and MCC separately for each speaker within each model, and then averaged the results across the ten models to obtain speaker-wise metrics. The results show that TLV recorded the lowest average FAR at 41.81%, followed by THV with the second lowest average FAR of 52.56%. Both speakers also demonstrated low average FRR (THV: 6.61%, TLV: 6.72%) and high average MCC scores, ranking first (0.56) and second (0.47) for TLV and THV, respectively.

5.2.3.3 Effect of utterance length

Figure 5.2.2 illustrates the relationship between utterance-wise DA and utterance length, measured by the number of phonemes or articulatory labels, across the ten models. A noteworthy overall trend is the superiority of the proposed models (M1 and M2) over their respective baselines (FP and FT). For example, PHN-M1 (dark purple curve) surpasses PHN-FP (dark yellow curve), highlighting the benefits of integrating AFs. Furthermore, a pattern consistent with the results described in Table 5.2.1 emerges when comparing models across the PHN and ART frameworks. Specifically, RS and FP achieve better results under the ART framework, whereas M1, FT, and M2 yield marginally better performance within the PHN framework.

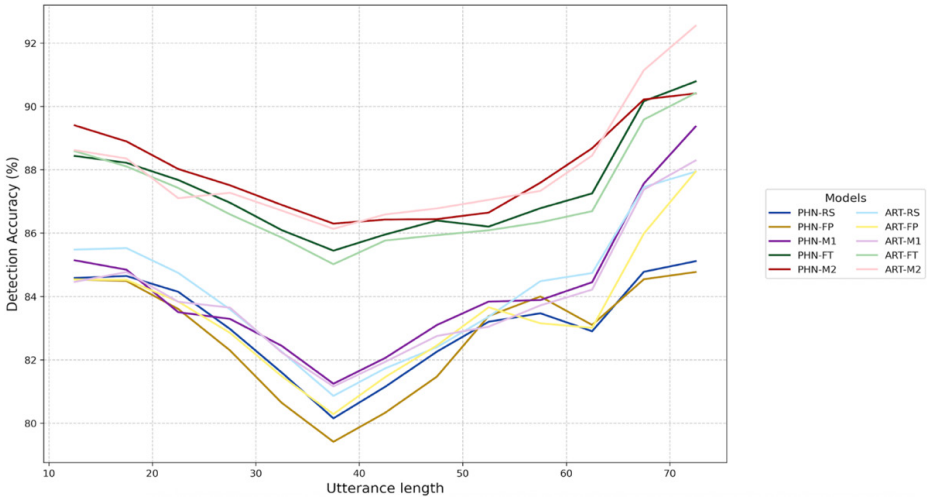

Figure 5.2.2: *Consistency rate of AFs in the L2 datasets. The horizontal axes denote the number of labels (phonemes/articulatory labels).*

As illustrated in Figure 5.2.2, model performance noticeably declines when processing utterances shorter than 40 labels, particularly in the 20-40 range. To better understand the factors contributing to this pattern, we first partitioned the test set into three utterance length categories: short (<21 labels), medium (21-40 labels), and long

(>40 labels). For each length category, FAR and FRR were computed separately for each of the ten models, resulting in three FAR and FRR values per model. These values were then averaged across the ten models within each length category to identify general trends. The test set is primarily composed of medium-length utterances, with fewer short and long cases. Our analysis reveals that the average FRR increases from 6.13% for short utterances to 8.55% for medium-length utterances, the highest among all categories, before slightly decreasing to 8.23% for long utterances. Similarly, the average FAR increases from 47.07% (short) to 53.32% (medium), with long utterances reaching 56.20%. These con-current increases in FAR and FRR account for the decline in DA from short to medium-length utterances, suggesting that models encounter greater challenges in handling intermediate-length utterances.

5.2.3.4 Diagnostic precision on frequent mispronunciations

A critical distinction between the PHN and ART frameworks, particularly when integrated with AFs, was observed in their diagnostic capabilities, as measured by DER. Figure 5.2.3 presents a heatmap illustrating the DER values associated with the ten most frequent mispronunciations made by the speakers in the test set, selected by ranking error types based on their frequency. The corresponding distributions of these ten mispronunciations are summarized in Table 5.2.4, where the "DH/D" substitution error emerges as the most frequent in the test set, followed by "Z/S", indicating their prevalence among the L2 speakers examined. Notably, these two error types also exhibit higher frequency in the entire dataset, which may partially explain their relatively lower DER values in Figure 5.2.3.

Table 5.2.4: *Number of the most frequent mispronunciations in the test set and their occurrences in the entire datasets in this study.*

Error type	Test set	Entire dataset
DH/D	348	1676
Z/S	296	1701
D/SIL	238	706
ER/AH	222	400
R/SIL	207	534
T/SIL	193	565
IH/IY	134	855
EY/EH	131	243
L/SIL	126	298
OW/AO	125	389

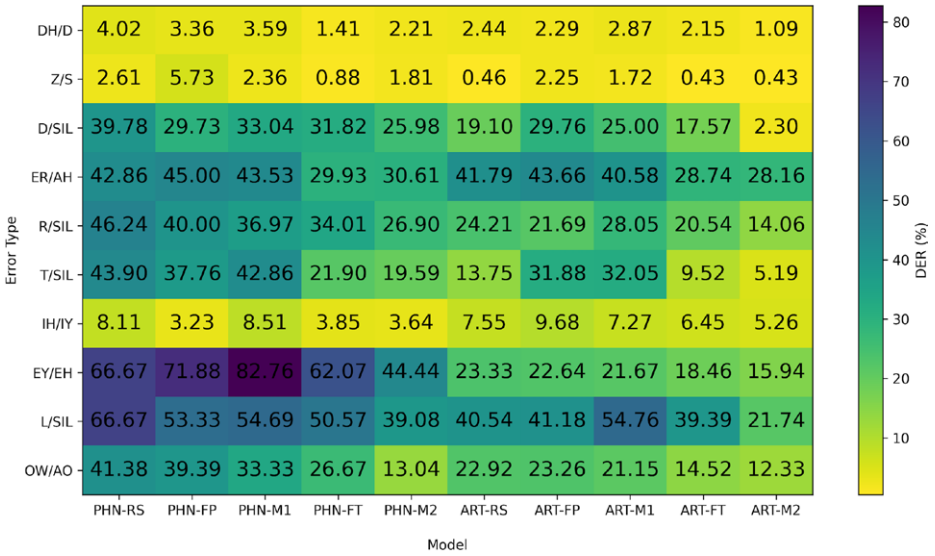

Figure 5.2.3: *DER for the most frequent mispronunciations in the test set. (Darker indicates higher DER).*

In Figure 5.2.3, the "DH/D" pair on the first row indicates that the phoneme "/DH/" was mispronounced as "/D/" by the speaker, representing a substitution error. As observed, the integration of AFs generally leads to reduced DER across both frameworks, especially for M2 vs. FT. More strikingly, ART models integrating AFs (ART-M1 and ART-M2) consistently obtained substantially lower DER values for most of these most frequent mispronunciations compared to the corresponding PHN models (PHN-M1 and PHN-M2) that integrated AFs. These findings highlight the superior capacity of the ART framework to leverage AFs for fine-grained, precise diagnosis of frequent mispronunciation types. However, despite these improvements, certain error types like L/SIL remain challenging, as evidenced by persistently high DER values even in the best-performing ART models, as detailed in Figure 5.2.3.

5.2.4 Discussion and conclusions

Through a systematic, multi-dimensional error analysis of AF-enhanced E2E MDD models, several key insights were gained that directly address our research questions. This in-depth examination of model behavior helped identify the core factors affecting both detection accuracy and diagnostic precision.

While the integration of AFs generally improved model performance across PHN and ART frameworks, our analysis for RQ5.2.1 also uncovered certain limitations and performance bottlenecks in current models. As shown in Figure 5.2.2, utterance-wise DA declines for utterances shorter than 40 labels. Note that medium-

length utterances make up approximately 65% of the test set. As demonstrated in Section 5.2.3.3, this length group also exhibits the highest FAR and FRR, indicating that models struggle the most in this region. Interestingly, the proportion of mispronunciations in the short, medium, and long utterances is relatively similar (16.76%, 16.95%, and 18.31%, respectively), suggesting that the performance drop is not simply due to error frequency. We hypothesize that medium-length utterances fall into a contextual 'gray zone': they are too long for simple modeling yet lack the redundancy for longer utterances needed for disambiguation. This limitation may be attributed to the insufficient receptive fields of current architectures, such as Conformer and XLSR. Similar observations have been reported in Cumbal et al. (2021), underscoring the need for improved context modeling, especially for mid-length speech segments.

Another salient finding related to RQ5.2.1 was the considerable inter-speaker variability noted, although the generalizability of this finding is constrained by the six-speaker sample. The contrast between THV and TLV serves as an example: THV showed the lowest DA despite its second-best average FAR and MCC, as detailed in Section 5.2.3.2, likely a consequence of its substantially higher error rate (27.00%), illustrated in Table 5.2.3, amplified false acceptances. Conversely, TLV's similar error rate resulted in higher DA, potentially due to more detectable errors (lowest average FAR: 41.81%). This implies that for L2 learners with high mispronunciation rates, detectability may be impacted by both frequency and distinctiveness. This remains a persistent challenge, even with AFs, and calls for validation on larger and more diverse L2 cohorts with varied L1 backgrounds.

Regarding the impact of output representations on the trade-off between DA and diagnostic precision (RQ5.2.2), the choice of model framework emerged as a key differentiator. While AFs generally enhanced performance over baselines, the nature of these gains differed across frameworks. ART models exhibited higher diagnostic precision, reflected in lower DER for common errors, whereas PHN models achieved comparable or slightly better DA. This suggests potential performance biases: ART outputs may better support error type identification, whereas PHN outputs tend to preserve phonetic detail, enhancing DA. These trends were further shaped by input configurations. RS and FP models performed better with ART outputs, likely due to their simplified classification categories. In contrast, M1 and M2 yield higher accuracy in the PHN setting, where the fine-grained AF inputs align more closely with phoneme outputs, maximizing the advantage of AF information. This benefit may diminish under ART, where coarser output granularity creates a mismatch with detailed inputs. Phoneme distribution may also contribute to these differences, with consonants accounting for 60.8% of L2-ARCTIC (Zhao et al., 2018), their

disproportionate influence may skew overall DA. Finally, PHN-FT outperformed ART-FT, possibly due to greater compatibility with the phoneme-centric pretraining of XLSR, facilitating more efficient fine-tuning.

In summary, our multi-dimensional error analysis identifies three principal findings for AF-enhanced MDD models: 1) The ART framework provides better diagnostic precision for most frequent errors compared to PHN models, despite marginally lower DA and MCC scores; 2) Considerable inter-speaker variability in DA is observed within the limited speakers, even with comparable proportions of mispronunciations, suggests differences in error detectability, potentially linked to acoustic-phonetic distinctiveness, may better explain model performance variations than error quantity alone; 3) While current AF-enhanced models show consistent improvements, they still struggle with medium-length utterances. These insights suggest that AF integration yields distinct but context-dependent L2 assessment advantages, highlighting intricate links among representation granularity, feature compatibility, and training strategies, motivating further studies on larger, diverse datasets.

This study has limitations, most notably the small speaker sample (N = 6), reducing statistical power and generalizability. While our multi-metric analysis revealed meaningful individual performance differences, a thorough investigation of acoustic-phonetic factors was beyond the scope of this work and needs to be addressed in future exploration. Future studies should: 1) validate the findings on larger and more diverse L2 datasets; 2) improve contextual modeling to better address utterance length; and 3) examine mispronunciation patterns at the acoustic-phonetic level.


Hello
你好

Chapter 6

Discussion and Conclusions

The primary objective of the research reported on in this thesis was to study how the automatic assessment of L2 speech can be enhanced by examining both speech intelligibility assessment and automatic pronunciation assessment, with a special focus on the role of acoustic-phonetic features, strategies for overcoming data scarcity in non-dominant pluricentric language varieties, and improving state-of-the-art (SOTA) modeling approaches with articulatory information. This chapter combines the findings from the empirical studies presented in **Chapters 2 to 5** and discusses them in relation to the original research questions in **Chapter 1**. Section 6.1 addresses these questions step by step, reflecting the order of the empirical chapters. For each question, the discussion reviews the scope of investigation and summarizes the main findings. Section 6.2 then offers a general conclusion that unites the contributions of all studies to highlight the broader implications of the thesis. Finally, Section 6.3 considers the limitations of the present work and outlines potential directions for future research.

6.1 Discussing the results

This section revisits the four main research questions outlined in Section 1.3. The discussion is organized by theme, reflecting the order of the empirical chapters, which begin with acoustic feature analysis and intelligibility measurement, and progress to system-level optimization for pluricentric languages and articulatory-enhanced modeling.

6.1.1 Acoustic-phonetic characteristics of non-native speech

The first research question (RQ1) aimed to identify which acoustic-phonetic features are most effective in distinguishing native from non-native speech. This inquiry was addressed in **Chapter 2** through a systematic comparison of large-scale acoustic-phonetic features using Support Vector Machines (SVM) and Recursive Feature Elimination (RFE). The statistical analyses showed that the acoustic differences between native and non-native speakers are extensive rather than limited to specific aspects. The results showed that the vast majority of the examined features demonstrated statistically significant differences between the two groups. This widespread distinctiveness confirms that non-native speech production fundamentally differs from native norms across multiple acoustic domains.

A key finding regarding specific signal characteristics is the strong discriminative power of loudness and articulation rate measures. The analysis of effect sizes demonstrated that non-native speakers consistently produced speech with notably

lower loudness levels compared to native speakers. This acoustic decrease supports the study's hypothesis that the lower linguistic confidence inherent to L2 production physically appears as reduced speaking volume. Likewise, articulation rate emerged as a reliable discriminator across different language backgrounds. The non-native group exhibits significantly lower mean articulation rate values as well as smaller standard deviations (SDs) compared to the native group, regardless of their first language. These findings indicate that overall differences in speaking pace and vocal energy serve as important acoustic indicators that characterize non-native speech production.

Finally, a discrepancy was noticed between feature rankings derived from RFE and effect size, where features with a high effect size were not always given priority by RFE. The investigation suggests that this difference may be due to the different aims of the two methods. While effect size evaluates the importance of features individually, RFE considers the dependencies and potential redundancies among features.

6.1.2 Intelligibility measurement and prediction

Building on the feature analysis in **Chapter 2**, the second research question (RQ2) examined how these features correlate with and predict the intelligibility of non-native speech in **Chapter 3**. This was addressed through a two-step investigation: first, by confirming the reliability of human measurement constructs in Section 3.1, and second, by developing automated predictive models based on these validated measures in Section 3.2.

The initial investigation into measurement reliability revealed that both Visual Analogue Scales (VAS) and Orthographic Transcription (OT)-based measures are reliable instruments for assessing non-native speech intelligibility, as long as they are used within their appropriate contexts. The statistical analyses demonstrated high internal consistency for both measures; however, a weak correlation was observed between them. This difference suggests that they measure different aspects of communication. It seems that VAS scores capture the intuitive and perceived magnitude of understanding, which can be interpreted as a subjective dimension related to speech intelligibility and is closely aligned with the notion of comprehensibility; while OT-based measures serve as objective indicators of functional intelligibility, demonstrating particular utility in identifying specific pronunciation problems. Taken together, these findings suggest that the choice of a ground truth metric should depend on the specific assessment goal.

Building on these validated constructs, the second stage of the investigation demonstrated that these human measures can be appropriately predicted using

automatically extracted acoustic-phonetic features. The regression analyses showed that validated intelligibility scores could be effectively modeled with a data-driven selection of acoustic-phonetic parameters. By applying dimensionality reduction techniques to identify the most relevant predictions from a large set of features, the proposed models successfully established a mapping between the acoustic signals and the corresponding human ratings. This confirms that speech intelligibility measurements, whether based on subjective ratings (VAS) or objective transcriptions (OTs), can be reliably modeled using quantifiable signal characteristics and thus effectively automated through the statistical selection of key features.

6.1.3 Data strategies for pluricentric varieties

The third research question (RQ3) addressed the challenge of data scarcity by exploring how cumulated speech resources from a dominant variety affect system performance for a non-dominant variety. **Chapter 4** investigates this question using Dutch as a case study, with Flemish Dutch as the target non-dominant variety and Netherlandic Dutch as the dominant source variety. The results show that, for both ASR and PED systems targeting Flemish Dutch, cross-variety resource sharing is not uniformly beneficial. Instead, its effectiveness depends on the specific computational task (ASR vs. PED) and the nature of the cumulated training data.

Regarding ASR, the results confirm that cumulating speech resources from the dominant variety is an effective strategy to improve performance on the non-dominant target variety. The experiments showed that combining Netherlandic and Flemish datasets for Flemish Dutch recognition generally led to lower Word Error Rates (WERs) compared to the baseline models trained only on limited target data (for example, the model Con-N1-F1, which was trained on a combination of Netherlandic and Flemish Dutch native speech, outperformed both Con-N1 and Con-F1, which were trained solely on native Netherlandic and Flemish Dutch speech, respectively). Notably, the greatest improvements for Flemish Dutch recognition models were observed in the model that used non-native speech resources from both varieties. This suggests that the ASR system benefits significantly from a larger amount of domain-matched learner data, which helps with generalization across interlanguage patterns. However, the study also points out that simply increasing data volume does not always lead to proportional improvements; the model trained on the full dataset (combining all native and non-native training data, Con-all) performed slightly worse than the model trained on only the combined non-native training data (Con-N2-F2). As discussed in **Chapter 4**, this indicates that while adding more data is helpful, the benefits can be limited by severe data imbalance and the extra complexity caused by diverse L1 backgrounds within the dominant non-native dataset. Additionally,

although transfer learning effectively leveraged knowledge from the dominant variety, it yielded less impressive results compared to the direct data cumulation approach, reinforcing the idea that for ASR, directly integrating cross-variety resources is the superior approach.

Turning to PED, the investigation examined whether the same data cumulation strategy could be effectively applied to error detection tasks. Using artificial errors to simulate specific mispronunciations, the study found that GOP-based models trained exclusively on the target variety (Flemish) consistently outperformed those trained on combined cross-variety datasets. The inclusion of Netherlandic data failed to yield improvements; instead, it resulted in a noticeable performance decline. This outcome might be attributed to the phonological variation between the varieties, such as systematic differences in vowel realizations, which created a mismatch between the combined training data and the specific target phonology. More importantly, this contrast highlights a fundamental difference between ASR and PED tasks: while ASR systems are designed to be relatively tolerant to phonetic variability as long as lexical recognition is preserved; PED systems, on the other hand, should be strict, requiring fine-grained sensitivity to subtle, variety-specific phonetic deviations. Therefore, the results suggest that for the specific task of error detection, increasing the dataset size through cross-variety data cumulation reduces the discriminatory power required to identify subtle phonetic deviations.

6.1.4 Articulatory-enhanced MDD models

The final research question (RQ4) focused on whether articulatory features (AFs) can enhance the detection and diagnosis of mispronunciations, particularly within end-to-end (E2E) modeling architectures. This question was addressed in **Chapter 5** through a two-part investigation: developing AF-integrated E2E architectures in Section 5.1 and performing a multi-dimensional error analysis to evaluate their robustness in Section 5.2.

The combined findings from the modeling experiments and the error analysis provide a comprehensive affirmative answer to RQ4. First, regarding mispronunciation detection, the results demonstrate that enhancing E2E models with explicit articulatory features consistently improves detection accuracy (DA) compared to corresponding baselines (e.g., M1 vs. FP and M2 vs. FT). This confirms that AFs provide essential stability for accurately accepting or rejecting L2 phones. Second, in terms of diagnosis, the results emphasize the importance of the output framework. While the phoneme-based (PHN) framework is practical for binary detection, the articulatory-based (ART) framework is crucial for reducing the diagnostic error

rate (DER), which is essential for keeping learners engaged in ongoing studies. By translating the model's internal representations into practical, sub-segmental feedback, the AF-integrated ART framework directly meets the diagnostic needs for RQ4, bridging the gap between error detection and diagnosis.

Furthermore, the answer to RQ4 is refined by a closer examination of utterance-level effects, which reveals a systematic and non-linear relationship between utterance length and mispronunciation detection performance. In particular, medium-length utterances consistently constitute the most challenging condition across all model variants examined. This pattern is unlikely to reflect suboptimal model training, but rather aligns with the intrinsic properties of the sequence alignment mechanisms underlying E2E ASR-based architectures. Specifically, Connectionist Temporal Classification (CTC), which forms the basis of alignment in the evaluated models, does not learn explicit frame-level boundaries but instead marginalizes over all possible alignment paths using a soft alignment formulation (Graves et al., 2006). While this design is effective for unsegmented sequence learning, it has been shown to produce peaky and temporally uncertain alignments at the phoneme level. Critically, unlike long utterances where natural pauses often facilitate implicit segmentation and state resetting, medium-length utterances in L2 speech often exhibit continuous yet dysfluent articulation. In such conditions where contextual redundancy is limited and acoustic boundaries are blurred, recent work has demonstrated that alignment uncertainty can substantially degrade forced-alignment precision and downstream phonetic analyses (Huang et al., 2024). In parallel, ASR studies have reported that utterance length and length mismatch constitute influential performance factors, further indicating that sequence-level properties can systematically affect model robustness (Lin et al., 2023).

Taken together, these observations provide a principled explanation for the degradation observed in medium-length utterances. Since current mispronunciation detection systems are predominantly built upon ASR-based architectures, they inherently inherit the alignment sensitivities of their underlying recognition frameworks. Consequently, although AF integration substantially enhances the phonetic distinctiveness and interpretability of L2 speech representations, its practical effectiveness remains constrained by utterance-level alignment reliability and speaker-specific variability. This finding highlights an important structural limitation of current E2E mispronunciation detection systems and underscores the need to consider alignment-aware modeling strategies in future work.

6.2 General conclusions

The research reported on in this thesis investigated the acoustic-phonetic characterization, intelligibility assessment, and automated detection and diagnosis of non-native speech, also within the context of pluricentric languages. By integrating feature-level statistical analysis with system-level architectural optimization, the research offers a comprehensive understanding of how L2 speech deviations can be quantified, modeled, and leveraged for computer-assisted language learning applications. By systematically addressing the trajectory from signal-level attributes to pedagogical feedback, this work provides a more integrated perspective on automatic speech assessment than has been previously offered in the field. Based on the findings summarized in the previous section, three general conclusions can be drawn.

First, the research concludes that the acoustic distinction between native and non-native speech is pervasive and structurally complex, necessitating multivariate analytical approaches. The empirical evidence confirms that non-native speech production deviates from native norms across a broad spectrum of spectral and temporal dimensions. Theoretically, this research advances the acoustic-phonetic characterization of L2 speech by providing a systematic hierarchy of features that distinguish non-native speech, moving beyond the isolated variables typically explored in prior literature. Specifically, the observed tendencies of lower loudness and articulation rates align with the hypothesis that lower linguistic confidence is associated with acoustic reduction. The methodological contribution of this part lies in the comparative analysis of feature ranking techniques; the observed discrepancy between feature ranking derived from effect size and RFE highlights the strong interdependencies among these acoustic features. This finding explicitly challenges the practice of evaluating features in isolation and demonstrates that the distinction between native and non-native speech is more accurately captured when feature interactions and potential redundancies are considered. Building on this, our research determines that intelligibility is a multi-dimensional construct, for which these quantifiable characteristics serve as robust predictors of human intelligibility assessments. A key innovation of this work lies in the systematic comparison between intuitive measures (VAS) and direct measures (OTs), revealing that they capture distinct communicative dimensions. Furthermore, the research demonstrates the feasibility of automatically predicting these scores using acoustic-phonetic features, confirming that both subjective intelligibility ratings and transcription-based accuracy can be modeled objectively. This nuanced understanding advances the measurement framework for speech intelligibility assessment, implying that automated systems should be intentionally calibrated according to whether the

pedagogical goal is to evaluate global comprehensibility or to identify specific pronunciation problems. By clarifying the functional differences between these metrics and establishing their predictability, the dissertation provides a more principled basis for designing and implementing automated speech intelligibility evaluation protocols.

Second, regarding the challenge of data scarcity in pluricentric languages, the thesis concludes that cross-variety data cumulation is a task-dependent strategy rather than a universal solution. The divergent outcomes between ASR and PED demonstrate that these technologies have conflicting data requirements. This insight provides a nuanced empirical contribution by demonstrating that the common strategy of cumulating cross-variety data, while beneficial for general recognition tasks, can be counterproductive for tasks requiring high phonetic precision like PED. For ASR, where the primary objective is robust recognition despite acoustic variability, the system benefits primarily from the increased statistical coverage provided by combining domain-matched learner data even when sourced from a different variety. In contrast, PED functions as a task that requires precise phonetic discrimination, where sensitivity to acoustic deviation is paramount. The findings establish that introducing cross-variety data into PED models creates a phonological mismatch that diminishes the discriminatory power required to detect specific phonetic deviations. Therefore, optimal system design in low-resource pluricentric contexts should weigh the benefits of increased data volume against the risks of acoustic mismatch. This involves prioritizing data quantity for recognition tasks and phonological specificity for error detection. These findings offer a critical gap in the literature regarding the robustness of CALL tools for low-resourced language varieties and provide a practical blueprint for optimizing system design in pluricentric contexts.

Third, the research demonstrates that integrating domain-specific articulatory knowledge into E2E architectures significantly enhances both the robustness and the diagnostic utility of mispronunciation detection systems. This integration offers a more interpretable framework for E2E modeling by introducing stable linguistic constraints. The experiments confirm that explicit AFs provide stable linguistic constraints that improve detection accuracy compared to standard acoustic baselines. More importantly, the transition from phoneme-based to articulatory-based frameworks addresses the critical gap between detection and diagnosis. By enabling the system to output sub-segmental feedback, this approach moves beyond binary correctness judgments to identifying the underlying production errors. However, the demonstrated sensitivity of these models to utterance length and inter-speaker variability suggests that while AF integration is theoretically sound and

practically effective, its implementation requires careful consideration of the input context to maintain stability. While the use of articulatory feature sis not unique in itself, the dissertation's contribution lies in the systematic multi-dimensional error analysis of these models, which provides a deeper understanding of how such linguistic information affects performance across different speaker and utterance context, thereby opening clearer avenues for future research into interpretable diagnostic systems.

6.3 Limitations and future directions

Although this thesis provides a comprehensive framework for the multidimensional analysis and automated modeling of non-native speech, several limitations inherent in the study design and methodology point towards critical avenues for future research.

Regarding the acoustic analysis and intelligibility assessment (RQ1 and RQ2), a primary limitation lies in the reliance on conventional, expert-defined acoustic features and global statistical functionals. While effective for identifying key acoustic indicators like loudness and articulation rate using classical models like SVMs, these aggregated metrics may not fully capture the complex, non-linear temporal dynamics of L2 speech as well as modern deep representation learning. Future research should explore Self-Supervised Learning (SSL) frameworks, such as Wav2Vec 2.0 or HuBERT, to automatically extract rich, contextualized acoustic representations directly from raw waveforms. Additionally, although the distinction between impression-based (VAS) and accuracy-based (OT) metrics has been established, current automated models primarily provide numerical predictions without accompanying educational explanations. To bridge the gap between assessment and learning, future systems can incorporate Explainable AI (XAI) techniques, leveraging methods like feature attribution or attention visualization to explain the decision-making process of "black box" models. This would enable models to not only predict an intelligibility score, but also pinpoint which specific acoustic deviations (e.g., "speaking rate too slow") most influence that score, thereby offering actionable feedback aligned with the specific metric used.

In the domain of data strategies for pluricentric varieties (RQ3), the limitations of the traditional cumulative approach became evident through the decline in performance observed in the full-combination dataset. The study treated cumulated data mainly as a uniform resource, which led to interference from diverse L1 backgrounds and data imbalance. Future research should go beyond indiscriminate cumulation

and explore Active Learning (AL) or Data Selection (DS) algorithms. By employing criteria such as uncertainty sampling or density matching, these methods can intelligently identify and select only the samples from the dominant variety that are mathematically advantageous for the target variety, effectively filtering out confusing L1 patterns. Additionally, instead of simply combining datasets, Domain Adaptation (DA) techniques, such as adversarial training or feature alignment that minimize the distributional discrepancy between varieties, could allow for more effective knowledge transfer. This would specifically address spectral shifts (e.g., mapping Netherlandic to Flemish acoustic space) without the phonological mismatch issues that hampered the PED performance in this study.

Finally, regarding the articulatory-enhanced modeling (RQ4), the multi-dimensional analysis revealed that even with advanced architectures like Conformer or fine-tuned Wav2Vec 2.0, robustness remains constrained by input context and speaker characteristics. Specifically, the models were sensitive to utterance length and inter-speaker variability, indicating that the current method of integrating AFs may not fully leverage their potential in dynamic contexts. Future work should address this by exploring Hierarchical Attention Mechanisms or Cross-modal Adaptive Fusion. Unlike simple feature concatenation, these more advanced methods would enable the model to adaptively weigh the importance of articulatory versus acoustic cues at different time scales, thereby boosting stability across various utterance lengths. Additionally, the current system depends on forced alignment to obtain ground-truth articulatory labels, which acts as a bottleneck when processing spontaneous speech. A promising direction is to explore joint optimization or fully E2E approaches, where articulatory representations are learned as latent variables, thereby reducing the reliance on external alignment tools and making the system more robust for real-world, interactive CALL applications.

6


Hello
你好

Appendices

References

List of Abbreviations

Research Data Management

English Summary

Nederlandse Samenvatting

Acknowledgements

Curriculum Vitae

List of publications

References

Abur, D., Enos, N. M., & Stepp, C. E. (2019). Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in Parkinson's disease with variable listener exposure. *American Journal of Speech-Language Pathology*, *28*(3), pp. 1222-1232.

Aoyama, K., Hong, L., Flege, J. E., Akahane-Yamada, R., & Yamada, T. (2023). Relationships Between Acoustic Characteristics and Intelligibility Scores: A Reanalysis of Japanese Speakers' Productions of American English Liquids. *Language and Speech*, *66*(4), 1030-1045.

Arai, K., Araki, S., Ogawa, A., Kinoshita, K., Nakatani, T., & Irino, T. (2020). Predicting Intelligibility of Enhanced Speech Using Posteriors Derived from DNN-Based ASR System. *Proc. Interspeech 2020*, pp. 1156-1160.

Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44-58.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Qiantong, L., Sriram, A., ... & Baevski, A. (2021). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint arXiv:2111.09296*.

Baese-Berk, M. M., & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *The Journal of the Acoustical Society of America*, *138*(3), EL223-EL228.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

Bai, Y., Hubers, F., Cucchiarini, C., & Strik, H. (2020). ASR-Based Evaluation and Feedback for Individualized Reading Practice. *Proc. Interspeech 2020*, pp. 3870-3874.

Baird, A., Amiriparian, S., Milling, M., & Schuller, B. W. (2021). Emotion recognition in public speaking scenarios utilising an LSTM-RNN approach with attention. In *2021 IEEE spoken language technology workshop (SLT)*, pp. 397-402.

Bälter, O., Engwall, O., Öster, A. M., & Kjellström, H. (2005). Wizard-of-Oz test of ARTUR: a computer-based speech training system with articulation correction. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility* (pp. 36-43).

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600-1610.

Bent, T., Bradlow, A. R., & Smith, B. L. (2008). Production and perception of temporal patterns in native and non-native speech. *Phonetica*, *65*(3), 131-147.

Benway, N. R., Siriwardena, Y. M., Preston, J. L., Hitchcock, E., McAllister, T., & Espy-Wilson, C. (2023). Acoustic-to-Articulatory Speech Inversion Features for Mispronunciation Detection of/r/in Child Speech Sound Disorders. *Proc.Interspeech 2023*, 4568-4572.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11), 763–786.

Best, C. T., Tyler, M., Bohn, O., & Munro, M. (2007). Nonnative and second-language speech perception. *Language experience in second language speech learning*, *17*, 13-34.

Bhalla, D. (2015). *Simplest Dimensionality Reduction with R*. ListenData. Retrieved from https://www.listendata.com/2015/06/simplest-dimensionality-reduction-with-r.html

Biadsy, F., Hirschberg, J., & Ellis, D. P. W. (2011). Dialect and accent recognition using phonetic-segmentation supervectors. *Proc. Interspeech 2011*, pp. 745–748.

Boersma, P. (2011). Praat: doing phonetics by computer [Computer program]. *http://www.praat.org/*.

Bohn, O. S., & Flege, J. E. (1992). The production of new and similar vowels by adult German learners of English. *Studies in second language acquisition*, *14*(2), 131-158.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3-4), 255-272.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155-180.

Burgos, P., Jani, M., Cucchiarini, C., van Hout, R., & Strik, H. (2014). Dutch vowel production by Spanish learners: duration and spectral features. Proc. *Interspeech 2014*, pp. 529-533.

Burleson, D. F. (2007). Improving intelligibility of non-native speech with computer-assisted phonological training. *IULC Working Papers*, *7*(1).

Bygate, M. (1987). *Speaking*. Oxford University Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1-47.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation hardback with audio CDs (2): A course book and reference guide*. Cambridge University Press.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4960-4964.

Chapelle, C. A. (2001). *Computer applications in second language acquisition*. Cambridge University Press.

Chen, N. F., & Li, H. (2016). Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1-7.

Chen, N. F., Wee, D., Tong, R., Ma, B., & Li, H. (2016). Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL. *Speech Communication*, *84*, 46-56.

Chen, Q., Lin, B., & Xie, Y. (2022). An Alignment Method Leveraging Articulatory Features for Mispronunciation Detection and Diagnosis in L2 English. *Proc. Interspeech 2022*, 4342-4346.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*, *28*.

Chun, D. M., & Levis, J. M. (2020). Prosody in second language teaching: Methodologies and effectiveness. *The Oxford Handbook of Language Prosody*, 618-630.

Chung, W. J., & Kang, H. G. (2024). Speaker-Independent Acoustic-to-Articulatory Inversion through Multi-Channel Attention Discriminator. *Proc. Interspeech 2024*, 1540-1544.

Clyne, M. (1992). *Pluricentric languages: Differing norms in different nations*. Mouton de Gruyter.

Clyne, M. (1997). Pluricentric languages and national identity: An antipodean view. *Englishes around the World*, pp. 287–300.

Colpaert, J. (2024). On the rationale behind Language and Motivation. *Language and Motivation*, 1(1), 1–4.

Colpaert, J., & Stockwell, G. (2022). *Smart CALL: Personalization, contextualization, & socialization*. Castledown Publishers.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Cook, V. (2016). *Second language learning and language teaching*. Routledge.

Crabtree, I. (2021). *The role of listener ideology in perception of non-native speech volume*. University of Oregon.

Cronk, B. C. (2016). *How to use IBM SPSS statistics: A step-by-step guide to analysis and interpretation*. Routledge.

Crystal, D. (2003). *English as a global language*. Cambridge university press.

Cucchiarini, C., Driesen, J., Hamme, H. V., & Sanders, E. P. (2008). Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus. *LREC 2008*.

Cucchiarini, C., Neri, A., & Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication, 51*(10), 853-863.

Cucchiarini, C., & Strik, H. (2017). Automatic speech recognition for second language pronunciation training. *The Routledge handbook of contemporary English pronunciation*, pp. 556-569. Routledge.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America, 107*(2), 989-999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and-spontaneous speech. *The Journal of the Acoustical Society of America, 111*(6), 2862-2873.

Cumbal, R., Moell, B., Lopes, J., & Engwall, O. (2021). You don't understand me!: comparing ASR results for L1 and L2 speakers of Swedish. *Proceedings of Interspeech 2021*, pp. 4463-4467.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(1), 30-42.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing, 28*(4), 357-366.

Deepanshu Bhalla. (2022). Dimensionality reduction with R. https://www.listendata.com/2015/06/simplest-dimensionality-reduction-with-r.html.

Despres, J., Fousek, P., Gauvain, J. L., Gay, S., Josse, Y., Lamel, L., & Messaoudi, A. (2009, September). Modeling northern and southern varieties of dutch for STT. *Proc. Interspeech 2009*, pp. 96-99.

Derakhshan, A., & Karimi, E. (2015). The interference of first language and second language acquisition. *Theory and Practice in language studies*, *5*(10), 2112-2117.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1-16.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379-397.

Dong, W., Cucchiarini, C., & Strik, H. (2023). End-to-End Mispronunciation Detection and Diagnosis for Non-native English Speech. *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 61–65.

Duan, R., Kawahara, T., Dantsuji, M., & Nanjo, H. (2019). Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 391-401.

Duan, R., Kawahara, T., Dantsujii, M., & Zhang, J. (2016). Pronunciation error detection using DNN articulatory model based on multi-lingual and multi-task learning. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1-5.

Duan, R., Kawahara, T., Dantsuji, M., & Zhang, J. (2017). Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning. *IEICE TRANSACTIONS on Information and Systems*, *100*(9), 2174-2182.

El Kheir, Y., Chowdhury, S. A., & Ali, A. (2024). L1-aware multilingual mispronunciation detection framework. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 12752-12756). IEEE.

Elffers, C., Van Bael, C., & Strik, H. (2005). *ADAPT: algorithm for dynamic alignment of phonetic transcriptions. Nijmegen: Department of Language & Speech, Radboud University*.

Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal: An Open Access Refereed Journal for World Language Educators*, *1*(1).

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology, 2*(2), 62-76.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication, 51*(10), 832-844.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 835-838).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459-1462.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, *7*(2), 190-202.

Fewzee, P., & Karray, F. (2012). Dimensionality reduction for emotional speech recognition. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pp. 532-537.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399-423.

Finizia, C., Lindström, J., & Dotevall, H. (1998). Intelligibility and perceptual ratings after treatment for laryngeal cancer: laryngectomy versus radiotherapy. *The Laryngoscope*, *108*(1), 138-143.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of phonetics*, *15*(1), 47-65.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, *92*(1), 233-277.

Fontan, L., Ferrané, I., Farinas, J., Pinquier, J., Tardieu, J., Magnen, C., Gaillard, P., Aumont, X., & Füllgrabe, C. (2017). Automatic Speech Recognition Predicts Speech Intelligibility and Comprehension for Listeners With Simulated Age-Related Hearing Loss. *Journal of Speech, Language, and Hearing Research*, 60(9), 2394–2405.

Franco, H., Bratt, H., Rossier, R., Rao Gadde, V., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, *27*(3), 401-418.

Franco, H., Neumeyer, L., Ramos, M., & Bratt, H. (1999). Automatic detection of phone-level mispronunciation for language learning. In *Eurospeech* (Vol. 99, pp. 851-854).

Frankel, J., Magimai-Doss, M., King, S., Livescu, K., & Çetin, O. (2007). Articulatory feature classifiers trained on 2000 hours of telephone speech. *Proc. Interspeech 2007*, pp. 2485–2488.

Gallardo, L. F., Möller, S., & Beerends, J. (2017). Predicting automatic speech recognition performance over communication channels from instrumental speech quality and intelligibility scores. In *INTERSPEECH* (pp. 2939-2943).

Ganzeboom, M. S., Bakker, M., Cucchiarini, C., & Strik, H. (2016). Intelligibility of disordered speech: Global and detailed scores.

Gao, L., Tejedor-Garcia, C., Strik, H., & Cucchiarini, C. (2024). Reading miscue detection in primary school through automatic speech recognition. *Proc. Interspeech 2024*, pp. 5153-5157.

Gao, Y., Xie, Y., Cao, W., & Zhang, J. (2015). A study on robust detection of pronunciation erroneous tendency based on deep neural network. *Proc. Interspeech 2015*, 693-696.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. NIST.

Gass, S. M., Behney, J., & Plonsky, L. (2020). *Second language acquisition: An introductory course*. Routledge.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*(1), 65-89.

Ghosh, P. K., & Narayanan, S. (2011). Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, *130*(4), EL251-EL257.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369-376.

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Wu, Y. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. *Proc. Interspeech 2020*, pp. 5036-5040.

Gutz, S. E., Stipancic, K. L., Yunusova, Y., Berry, J. D., & Green, J. R. (2022). Validity of off-the-shelf automatic speech recognition for assessing speech intelligibility and speech severity in speakers with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, *65*(6), 2128-2143.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389-422.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL quarterly*, *38*(2), 201-223.

Hajiramezanali, E., Zamani Dadaneh, S., Karbalayghareh, A., Zhou, M., & Qian, X. (2018). Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. *Advances in Neural Information Processing Systems*, *31*.

Han, Z. H. (2004). *Fossilization in adult second language acquisition*. Multilingual Matters.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, *116*(5), 3108-3118.

Hazen, T. J., Shen, W., & White, C. (2009). Query-by-example spoken term detection using phonetic posteriorgram templates. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 421-426). IEEE.

Hecht, B. F., & Mulford, R. (1982). The acquisition of a second language phonology: Interaction of transfer and developmental factors. *Applied Psycholinguistics*, *3*(4), 313-328.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*(4), 1738-1752.

Hiligsmann, P., & Rasier, L. (2006). Uitspraakleer Nederlands voor Franstaligen.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, *29*(6), 82-97.

Höhle, B., & Weissenborn, J. (2003). German-learning infants' ability to detect unstressed closed-class elements in continuous speech. *Developmental Science*, *6*(2), 122-127.

Hu, W., Qian, Y., & Soong, F. K. (2015a). An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech. *SLaTE*, *5*, 71-76.

Hu, W., Qian, Y., Soong, F. K., & Wang, Y. (2015b). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, *67*, 154-166.

Huang, H., Xu, H., Hu, Y., & Zhou, G. (2017). A transfer learning approach to goodness of pronunciation based automatic mispronunciation detection. *The Journal of the Acoustical Society of America*, *142*(5), 3165-3177.

Huang, J. T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7304-7308.

Huang, R., Zhang, X., Ni, Z., Sun, L., Hira, M., Hwang, J., ... & Khudanpur, S. (2024). Less peaky and more accurate ctc forced alignment by label priors. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11831-11835). IEEE.

Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Hubbard, P. (2008). CALL and the future of language teacher education. *CALICO journal*, *25*(2), 175-188.

Hubers, F., Cucchiarini, C., & Strik, H. (2021). L2 idiom learning and L1-L2 similarity. *Proceedings of 12th International Conference of Experimental Linguistic 2021*.

Huijbregts, M., Ordelman, R., van der Werff, L., & de Jong, F. M. (2009). SHoUT, the university of twente submission to the n-best 2008 speech recognition evaluation for dutch. Proc. *Interspeech 2009*, pp. 2575-2578.

Inceoglu, S., Chen, W. H., & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL*, *35*(1), 89-104.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on L2 listeners' scalar ratings. *Studies in Second Language Acquisition*, *34*(3), 475-505.

Ito, A., Nagasawa, T., Ogasawara, H., Suzuki, M., & Makino, S. (2006). Automatic detection of English mispronunciation using speaker adaptation and automatic assessment of English intonation and rhythm. *Educational technology research*, *29*(1-2), 13-23.

Jack, R., & Richards, M. (2008). Teaching listening and speaking from theory to practice. *Cambridge University.*

Jackson, C. N., & O'Brien, M. G. (2011). The Interaction between Prosody and Meaning in Second Language Speech Production 1. *Die Unterrichtspraxis/Teaching German*, *44*(1), 1-11.

Jaitly, N., & Hinton, G. E. (2013). Vocal tract length perturbation (VTLP) improves speech recognition. *Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Janbakhshi, P., Kodrasi, I., & Bourlard, H. (2019). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6405-6409.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Joshi, S., Deo, N., & Rao, P. (2015). Vowel mispronunciation detection using DNN acoustic models with cross-lingual training. Proc. *Interspeech 2015*, 697-701.

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*(1), 115-146.

Kanters, S., Cucchiarini, C., Strik, H. (2009). The goodness of pronunciation algorithm: a detailed performance study. *Proc. Speech and Language Technology in Education (SLaTE 2009)*, 49-52.

Karbasi, M., & Kolossa, D. (2022). ASR-based speech intelligibility prediction: A review. *Hearing Research*, *426*, 108606.

Kearns, J. (2014). Librivox: Free public domain audiobooks.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review, 64*(3), 459-489.

Khanal, S., Johnson, M. T., & Bozorg, N. (2021, January). Articulatory comparison of l1 and l2 speech for mispronunciation diagnosis. In *2021 IEEE Spoken Language Technology Workshop (SLT)* (pp. 693-697). IEEE.

Kheir, Y. E., Ali, A., & Chowdhury, S. A. (2023). Automatic Pronunciation Assessment--A Review. *In Findings of the Association for Computational Linguistics: EMNLP 2023*, 8304-8324.

Kim, S. E., Chernyak, B. R., Seleznova, O., Keshet, J., Goldrick, M., & Bradlow, A. R. (2024). Automatic recognition of second language speech-in-noise. *JASA Express Letters*, *4*(2).

Kirchhoff, K., Fink, G. A., & Sagerer, G. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech communication, 37*(3-4), 303-319.

Kloots, H., Coussé, E., & Gillis, S. (2006). Vowel labelling in a pluricentric language: Flemish and Dutch labellers at work. *Linguistics in the Netherlands*, *23*(1), 126-136.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017, March). A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*,5220-5224.

Koffi, E. (2015). The pronunciation of voiceless th in seven varieties of l2 englishes: Focus on intelligibility. *Linguistic Portfolios*, *4*(1), 2.

Koffi, E. (2021). *Relevant acoustic phonetics of L2 English: Focus on intelligibility*. CRC Press.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145-164.

Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., & Stober, S. (2017). Transfer Learning for Speech Recognition on a Budget. *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP 2017 at the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, 168–177.

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press.

Lee, A., & Glass, J. (2012). A comparison-based approach to mispronunciation detection. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pp. 382-387.

Leung, K. K., Wu, M., & Meng, H. (2022). Contrastive Learning for Mispronunciation Detection. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Leung, W. K., Liu, X., & Meng, H. (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)*, pp. 8132-8136.

Levelt, W. J. (1993). *Speaking: From intention to articulation*. MIT press.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369-377.

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

Levy, M. (1997). *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Levy, M., & Stockwell, G. (2013). *CALL dimensions: Options and issues in computer-assisted language learning*. Routledge.

Li, K., Qian, X., & Meng, H. (2016a). Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(1), 193-207.

Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C. H. (2016b). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 6135-6139.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lin, J., Gao, Y., Zhang, W., Wei, L., Xie, Y., & Zhang, J. (2020). Improving pronunciation erroneous tendency detection with multi-model soft targets. *Journal of Signal Processing Systems*, *92*(8), 793-803.

Lin, Y. Y., Han, T., Xu, H., Pham, V. T., Khassanov, Y., Chong, T. Y., ... & Ma, Z. (2023). Random utterance concatenation based data augmentation for improving short-video speech recognition. *Proc. Interspeech 2023*, pp. 904-908.

Liu, P., Yu, Q., Wu, Z., Kang, S., Meng, H., & Cai, L. (2015). A deep recurrent approach for acoustic-to-articulatory inversion. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4450-4454.

Lo, T. H., Weng, S. Y., Chang, H. J., & Chen, B. (2020). An Effective End-to-End Modeling Approach for Mispronunciation Detection. *Proc. Interspeech 2020*, pp. 3027-3031.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Lybaert, C., & Delarue, S. (2017). Stereotypes and attitudes in a pluricentric language area. The case of Belgian Dutch. *Stereotypes and linguistic prejudices in Europe*, *175*.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in second language acquisition*, *32*(2), 265-302.

Ma, J., Hu, Y., & Loizou, P. C. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, *125*(5), 3387-3405.

Mahr, T. J., Berisha, V., Kawabata, K., Liss, J., & Hustad, K. C. (2021). Performance of forced-alignment algorithms on children's speech. *Journal of Speech, Language, and Hearing Research*, *64*(6S), 2213-2222.

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Routledge.

Mao, S., Wu, Z., Li, R., Li, X., Meng, H., & Cai, L. (2018a). Applying multitask learning to acoustic-phonemic model for mispronunciation detection and diagnosis in l2 english speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6254-6258).

Mao, S., Wu, Z., Li, X., Li, R., Wu, X., & Meng, H. (2018b). Integrating articulatory features into acoustic-phonemic model for mispronunciation detection and diagnosis in l2 english speech. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 1-6.

Martinez, A. M. C., Spille, C., Roßbach, J., Kollmeier, B., & Meyer, B. T. (2022). Prediction of speech intelligibility with DNN-based performance measures. *Computer Speech & Language*, *74*, 101329.

Matassoni, M., Gretter, R., Falavigna, D., & Giuliani, D. (2018). Non-native children speech recognition through transfer learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6229-6233). IEEE.

Mathad, V. C., Mahr, T. J., Scherer, N., Chapman, K., Hustad, K. C., Liss, J., & Berisha, V. (2021). The Impact of Forced-Alignment Errors on Automatic Pronunciation Evaluation. *Proc. Interspeech 2021*, 1922-1926.

Mefferd, A. S., & Green, J. R. (2010). Articulatory-to-acoustic relations in response to speaking rate and loudness manipulations. *Journal of Speech, Language, and Hearing Research*, *53*(5), 1206-1219.

Mencke, E. O., Ochsner, G. J., & Testut, E. W. (1983). Listener judges and the speech intelligibility of deaf children. *Journal of Communication Disorders*, *16*(3), 175-180.

Meng, H. (2009). Developing speech recognition and synthesis technologies to support computer-aided pronunciation training for Chinese learners of English. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation* (pp. 40-42). Waseda University.

Meng, H., Lo, Y. Y., Wang, L., & Lau, W. Y. (2007). Deriving salient learners' mispronunciations from cross-language phonological comparisons. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 437-442.

Middag, C., Bocklet, T., Martens, J. P., & Nöth, E. (2011). Combining Phonological and Acoustic ASR-Free Features for Pathological Speech Intelligibility Assessment. *Proc. Interspeech 2011*, pp. 3005-3008.

Mohamed, A. R., Hinton, G., & Penn, G. (2012). Understanding how deep belief networks perform acoustic modelling. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4273-4276.

Morshed, M., & Hasegawa-Johnson, M. (2022). Cross-lingual articulatory feature information transfer for speech recognition using recurrent progressive neural networks. *Proceedings of Interspeech 2022*, pp. 2298-2302.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *45*(1), 73-97.

Munro, M. J., & Derwing, T. M. (2015). Intelligibility in research and practice: Teaching priorities. *The handbook of English pronunciation*, 375-396.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in second language acquisition*, *28*(1), 111-131.

Nazir, F., Majeed, M. N., Ghazanfar, M. A., & Maqsood, M. (2019). Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for Arabic phonemes. *IEEE Access*, *7*, 52589-52608.

Neri, A., Cucchiarini, C., & Strik, H. (2006). Selecting segmental errors in non-native Dutch for optimal pronunciation training. *IRAL - International Review of Applied Linguistics in Language Teaching*, 44(4), 357–404.

Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL*, *20*(2), 225-243.

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., ... & Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, *4*(2), 182-207.

Odijk, J. E. J. M. (2012). Overleeft het Nederlands het digitale tijdperk?. *Neerlandia/Nederlands van nu*, *116*(4), 32-33.

Oostdijk, N. (2002). The design of the spoken Dutch corpus. In *New frontiers of corpus research* (pp. 105-112). Brill.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206-5210.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *Proceedings of Interspeech 2019*, 2613-2617.

Park, S., & Culnan, J. (2019). A comparison between native and non-native speech for automatic speech recognition. *The Journal of the Acoustical Society of America*, *145*(3_Supplement), 1827-1827.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* (Vol. 1, pp. 5-1).

Prabhavalkar, P., Rao, K., Sainath, T. N., Li, B., Johnson, L., & Jaitly, N. (2017). A comparison of sequence-to-sequence models for speech recognition. *Proc. Interspeech 2017*.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492-28518). PMLR.

Ren, Y., Kim, S. S., Hasegawa-Johnson, M., & Cole, J. (2004). Speaker-independent automatic detection of pitch accent. In *Proc. Speech Prosody* (pp. 521-524).

Reves, T., & Medgyes, P. (1994). The non-native English speaking EFL/ESL teacher's self-image: An international survey. *System*, *22*(3), 353-367.

Riney, T., & Anderson-Hsieh, J. (1993). Japanese pronunciation of English. *JALT Journal*, *15*(1), 21-36.

Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., & Wu, Z. (2019). Speech recognition with augmented synthesized speech. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, 996-1002.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*(10.5555), 26853.

Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., ... & Kursun, O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE journal of biomedical and health informatics*, *17*(4), 828-834.

Samarakoon, L., Mak, B., & Lam, A. Y. (2018). Domain adaptation of end-to-end speech recognition in low-resource settings. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 382-388). IEEE.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *Proceedings of Interspeech*.

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. *Proceedings of Interspeech 2009*, 312-315.

Schultz, T., & Waibel, A. (1998). Multilingual and crosslingual speech recognition. In *Proc. DARPA Workshop on Broadcast News Transcription and Understanding* (pp. 259-262).

Seide, F., Li, G., & Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. Proc. *Interspeech 2011*, pp. 437-440.

Shahin, M., & Ahmed, B. (2024). Phonological-Level Mispronunciation Detection and Diagnosis. In *Proc. Interspeech 2024*, pp. 307-311.

Shekar, R. C., Yang, M., Hirschi, K., Looney, S., Kang, O., & Hansen, J. H. (2023). Assessment of non-native speech intelligibility using wav2vec2-based mispronunciation detection and multi-level goodness of pronunciation transformer. *Proc. Interspeech 2023*, 984-988.

Shi, Q., Li, K., Zhang, S., Chu, S. M., Xiao, J., & Ou, Z. (2010). Spoken English assessment system for non-native speakers using acoustic and prosodic features. *Proc. Interspeech 2010*, 1874-1877.

Shon, S., Ali, A., & Glass, J. (2018). Convolutional neural networks and language embeddings for end-to-end dialect recognition. *Speaker Odyssey 2018, The Speaker and Language Recognition Workshop*.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Siriwardena, Y.M., Attia, A.A., Sivaraman, G., & Espy-Wilson, C.Y. (2022). Audio Data Augmentation for Acoustic-to-Articulatory Speech Inversion. *2023 31st European Signal Processing Conference (EUSIPCO)*, 301-305.

Spille, C., Ewert, S. D., Kollmeier, B., & Meyer, B. T. (2018). Predicting speech intelligibility with deep neural networks. *Computer Speech & Language*, *48*, 51-66.

Spyns, P., & Odijk, J. (2013). Essential speech and language technology for Dutch. *Results by the STEVIN programme: Springer*, *10*, 978-3.

Sridhara, S., Ramesh, N., Gopakkali, P., Das, B., Venkatappa, S. D., Sanjivaiah, S. H., ... & Elansary, H. O. (2020). Weather-based neural network, stepwise linear and sparse regression approach for rabi sorghum yield forecasting of Karnataka, India. *Agronomy*, *10*(11), 1645.

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press.

Strik, H., Daelemans, W., Binnenpoorte, D. M., Sturm, J., Vriend, F. D., & Cucchiarini, C. (2002). Dutch HLT resources: From BLARK to priority lists.

Strik, H., Truong, K. P., Wet, F. D., & Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection. *Proc. Interspeech 2007*, 1837-1840.

Strik, H., Truong, K., De Wet, F., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech communication*, *51*(10), 845-852.

Sun, Y., Huang, Q., Wu, X., & Perception, M. (2022). Unsupervised Acoustic-to-Articulatory Inversion with Variable Vocal Tract Anatomy. *Proc. Interspeech 2022*, pp. 4656-4660.

Tafazoli, D., Huertas-Abril, C. A., & Gomez-Parra, M. E. (2019). Technology-based review on Computer-Assisted Language Learning: A chronological perspective. *Píxel-Bit. Revista de Medios y Educación*, (54), 29-43.

Tepperman, J. (2012). Automatic pronunciation assessment. In *The Oxford Handbook of Language Testing*. Oxford University Press.

Thomas, S., Ganapathy, S., & Hermansky, H. (2012). Multilingual MLP features for low-resource LVCSR systems. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4269-4272). IEEE.

Thomson, R. (2017). Measurement of accentedness, intelligibility, and comprehensibility. In *Assessment in second language pronunciation* (pp. 11-29). Routledge.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied linguistics*, *36*(3), 326-344.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Toivola, M., Lennes, M., & Aho, E. (2009). Speech rate and pauses in non-native Finnish. *Proc. Interspeech 2009*, pp. 1707-1710.

Toutios, A., & Margaritis, K. (2003). Acoustic-to-articulatory inversion of speech: A review. *Proceedings of the International 12th TAINN*.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, *33*(3), 339-372.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 1-30.

Truong, K. P., Neri, A., Cucchiarini, C., & Strik, H. (2004). Automatic pronunciation error detection: an acoustic-phonetic approach. *Proc. InSTIL/ICALL Symposium*, pp. 135-138.

Tu, M., Grabek, A., Liss, J., Berisha, V. (2018) Investigating the Role of L1 in Automatic Pronunciation Evaluation of L2 Speech. *Proc. Interspeech 2018*, 1636-1640.

van Bemmel, L., Harmsen, W., Cucchiarini, C., & Strik, H. (2021). Automatic selection of the most characterizing features for detecting COPD in speech. In *International Conference on Speech and Computer*, pp. 737-748.

van de Velde, H., Kissine, M., Tops, E., van der Harst, S., & van Hout, R. (2010). Will Dutch become Flemish? Autonomous developments in Belgian Dutch. *Multilingua, 29*(3-4), 385-416.

van der Harst, S., Van de Velde, H., & Van Hout, R. (2014). Variation in Standard Dutch vowels: The impact of formant measurement methods on identifying the speaker's regional origin. *Language Variation and Change*, *26*(2), 247-272.

van Doremalen, J., Cucchiarini, C., & Strik, H. (2013). Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch. *The Journal of the Acoustical Society of America*, *134*(2), 1336-1347.

van Doremalen, J., Strik, H., & Cucchiarini, C. (2010). Optimizing Automatic Speech Recognition for low-proficient non-native speakers. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*.

van Leeuwen, D. A. (2013). N-Best 2008: A Benchmark Evaluation for Large Vocabulary Speech Recognition in Dutch. *Essential Speech and Language Technology for Dutch*, pp. 271–288.

van Maastricht, L., Krahmer, E., & Swerts, M. (2016). Native speaker perceptions of (non-) native prominence patterns: Effects of deviance in pitch accent distributions on accentedness, comprehensibility, intelligibility, and nativeness. *Speech Communication*, *83*, 21-33.

van Santen, J. P., Prud'hommeaux, E. T., & Black, L. M. (2009). Automated assessment of prosody production. *Speech communication*, *51*(11), 1082-1097.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Wang, H. W., Yan, B. C., Chiu, H. S., Hsu, Y. C., & Chen, B. (2022a). Exploring non-autoregressive end-to-end neural modeling for English mispronunciation detection and diagnosis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6817-6821). IEEE.

Wang, J., Liu, J., Zhao, L., Wang, S., Yu, R., & Liu, L. (2022b). Acoustic-to-articulatory inversion based on speech decomposition and auxiliary feature. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4808-4812.

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language teaching, 31*(2), 57-71.

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1240-1253.

Wei, X., Chen, J., Wang, W., Xie, Y., & Zhang, J. (2017). A study of automatic annotation of PETs with articulatory features. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1608-1612). IEEE.

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102-127). University of Michigan Press.

Wewers, M. E., & Lowe, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in nursing & health, 13*(4), 227-236.

Wieling, M., Sivaraman, G., & Espy-Wilson, C. (2017). Analysis of acoustic-to-articulatory speech inversion across different accents and languages. *Proc. Interspeech 2017*, 974-978.

Willemyns, R. (2013). *Dutch: Biography of a language*. Oxford University Press.

Witt, S. M. (1999). *Use of Speech Recognition in Computer-assisted Language Learning*. Ph.D. dissertation, University of Cambridge, Cambridge, United Kingdom.

Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. In *International Symposium on automatic detection on errors in pronunciation training* (Vol. 1).

Witt, S. M., & Young, S. J. (1997). Language learning based on non-native speech recognition. In *Eurospeech* (pp. 633-636).

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication, 30*(2-3), 95-108.

Wu, M., Li, K., Leung, W. K., & Meng, H. (2021). Transformer Based End-to-End Mispronunciation Detection and Diagnosis. Proc. *Interspeech 2021*, 3954-3958.

Xie, X., Liu, X., & Wang, L. (2016). Deep Neural Network Based Acoustic-to-Articulatory Inversion Using Phone Sequence Information. *Proc. Interspeech 2016*, 1497-1501.

Xu, X., Kang, Y., Cao, S., Lin, B., & Ma, L. (2021). Explore wav2vec 2.0 for Mispronunciation Detection. Proc. *Interspeech* 2021, 4428-4432.

Xue, W., Cucchiarini, C., van Hout, R. W. N. M., & Strik, H. (2019). Acoustic correlates of speech intelligibility. the usability of the eGeMAPS feature set for atypical speech. *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 48-52.

Xue, W., Hout, R.v., Boogmans, F., Ganzeboom, M., Cucchiarini, C., Strik, H. (2021) Speech Intelligibility of Dysarthric Speech: Human Scores and Acoustic-Phonetic Features. *Proc. Interspeech 2021*, pp. 2911-2915.

Xue, W., Mendoza Ramos, V., Harmsen, W. N., Cucchiarini, C., Van Hout, R. W. N. M., & Strik, H. (2020). Towards a comprehensive assessment of speech intelligibility for pathological speech. *Proc. Interspeech 2020*, pp. 3146-3150.

Yan, B. C., Wang, H. W., Wang, Y. C., & Chen, B. (2023). Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

Yan, B. C., Wu, M. C., Hung, H. T., & Chen, B. (2020). An End-to-End Mispronunciation Detection System for L2 English Speech Leveraging Novel Anti-Phone Modeling. *Proc. Interspeech* 2020, pp. 3032-3036.

Yang, M., Hirschi, K., Looney, S. D., Kang, O., & Hansen, J. H. (2022). Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment. *Proc. Interspeech 2022, 4481-4485*.

Ye, W., Mao, S., Soong, F., Wu, W., Xia, Y., Tien, J., & Wu, Z. (2022). An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (APL) embeddings. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6827-6831). IEEE.

Yilmaz, E., Ganzeboom, M. S., Cucchiarini, C., & Strik, H. (2016). Combining non-pathological data of different language varieties to improve DNN-HMM performance on pathological speech. *Proc. Interspeech 2016*, pp. 218-222.

Yorkston, K. M., & Beukelman, D. R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of communication disorders*, *11*(6), 499-512.

Yorkston, K. M., Strand, E. A., & Kennedy, M. R. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology*, *5*(1), 55-66.

Yoon, S. Y., Hasegawa-Johnson, M., & Sproat, R. (2010). Landmark-based automated pronunciation error detection. *Proc. Interspeech 2010,* pp. 614-617.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... & Woodland, P. (2002). The HTK book. *Cambridge university engineering department*, *3*(175), 12.

Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., ... & Qian, Y. (2015). Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 338-345.

Zhang, F., Huang, C., Soong, F. K., Chu, M., & Wang, R. (2008). Automatic mispronunciation detection for Mandarin. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5077-5080). IEEE.

Zhang, L., Zhao, Z., Ma, C., Shan, L., Sun, H., Jiang, L., ... & Gao, C. (2020). End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture. *Sensors*, *20*(7), 1809.

Zhang, Z., Wang, Y., & Yang, J. (2022). End-to-end Mispronunciation Detection with Simulated Error Distance. *Proc. Interspeech 2022*, 4327-4331.

Zhao, G., Toth, A., & Black, A. (2018). L2-ARCTIC: A non-native English speech corpus. *Proceedings of Interspeech 2018*, 2783-2787.

Zhao, T., Hoshino, A., Suzuki, M., Minematsu, N., & Hirose, K. (2012). Automatic Chinese pronunciation error detection using SVM trained with structural features. In *2012 IEEE Spoken Language Technology Workshop (SLT)* (pp. 473-478).

Zheng, J., Huang, C., Chu, M., Soong, F. K., & Ye, W. P. (2007). Generalized segment posterior probability for automatic mandarin pronunciation evaluation. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (Vol. 4, pp. IV-201). IEEE.

Zhi, N., & Li, A. (2021). Acoustic salience in the evaluations of intelligibility and foreign accentedness of nonnative vowel production. *Lingua*, *256*, 103069.

Zhong, H., Xie, Y., & Yao, Z. (2024). Leveraging Large Language Models to Refine Automatic Feedback Generation at Articulatory Level in Computer Aided Pronunciation Training. In *Proc. Interspeech 2024*, pp. 2600-2604.

Zhou, Y., Jin, R., & Hoi, S. C. H. (2010). Exclusive lasso for multi-task feature selection. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 988-995). JMLR Workshop and Conference Proceedings.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320.

List of Abbreviations

AAI	Acoustic-to-articulatory inversion
ACs	Articulatory categories
AFs	Articulatory features
AM	Acoustic model
ART	Articulatory-based
ASR	Automatic speech recognition
BLSTM	Bidirectional long short-term memory
CALL	Computer assisted language learning
CAPT	Computer assisted pronunciation training
CGN	Corpus gesproken Nederlands (The spoken Dutch corpus)
COPD	Chronic obstructive pulmonary disease
CTC	Connectionist temporal classification
DA	Detection accuracy
DCT	Discrete cosine transform
DER	Diagnostic error rate
DNN	Deep neural network
eGeMAPS	Extended version of GeMAPS
E2E	End-to-end
EER	Equal error rate
EMA	Electromagnetic articulography
FAR	False acceptance rate
FBanks	Filterbank energies
FP	FBank pitch
FRR	False rejection rate
FT	Fine-tune
G2P	Grapheme to phoneme
GeMAPS	Geneva minimalistic acoustic parameter set
GMM	Gaussian mixture model
GOP	Goodness of pronunciation
HMM	Hidden Markov model
ICC	Intraclass correlation coefficient
LM	Language model
LR	linear regression
ISLA	Idiomatic second language acquisition
L1	First language
L2	Second language
LASSO	Least absolute shrinkage and selection operator

LOOCV	Leave one out cross-validation
LOSO	Leave-one-subject-out
MAE	Mean absolute error
MCC	Mathews correlation coefficient
MDD	Mispronunciation detection & diagnosis
MFCC	Mel-frequency cepstral coefficients
MLP	Multilayer perceptron
OTs	Orthographic transcriptions
PCA	Principal component analysis
PED	Pronunciation error detection
PER	Phoneme error rate
PHN	Phoneme-based
PLCLs	Pluricentric languages
PLP	Perceptual linear prediction
RFE	Recursive feature elimination
RMSE	Root mean squared error
ROLEG	Radboud online linguistic experiment generator
RS	Raw speech
RVR	Redundant variable removement
SD	Dtandard deviation
SI	Speech intelligibility
SLL	Second language learning
SLR	Stepwise linear regression
SSL	Self-supervised learning
SOTA	State-of-the-art
SVM	Support vector machines
SUS	Semantically unpredictable sentences
TL	Transfer learning
VAS	Visual analogue scale
VTLP	Vocal tract length perturbation
WER	Word error rate

Research Data Management

Existing and collected data

Except for chapters 1 and 6, existing and/or collected data were used in the research described in every other chapter. In Chapter 2, audio recordings from two existing corpora, the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN; Oostdijk et al., 2002) and JASMIN-CGN (Cucchiarini et al., 2008), were selected for analysis of distinctive features between native and non-native speech. Selected audio recordings from these two corpora were also used to train and adapt acoustic models in Chapter 4. Moreover, Dutch non-native idiom recordings were also selected from two personal datasets collected by Cucchiarini et al. (2020) and Ferdy et al. (2021) for comparative analysis.

In Chapter 3, data related to the judgments in the listening experiments were collected through an online platform, and personal information, including age, gender, and experience with language teaching, was collected. The age and gender data were collected to conduct potential studies about their influence on intelligibility assessments. The information about their experience with language teaching was used to demonstrate whether the participants were expert listeners, in contrast to non-expert listeners, who may result in different intelligibility results.

In Chapter 5, a published English native speech corpus (Librispeech, Panayotov et al., 2015) was included to train acoustic models for articulatory feature extraction, and another published English L2 corpus (L2-ARCTIC, Zhao et al., 2018) was used to train and test the acoustic models for mispronunciation detection tasks.

Informed consent

For the listening experiments conducted to assess intelligibility in Chapter 3, human experts, all of whom were Dutch as a second language (NT2) teachers, were recruited through a private Facebook group that was not accessible to the public. The listening experiments were executed using the Radboud Online Linguistic Experiment Generator (ROLEG) platform. At the beginning of the experiments, the participants received information regarding the handling of their data, including details such as age, gender, and their experience in NT2 education. They were also provided with a consent form that they were required to sign. This process adhered to the informed consent procedure established by the Ethics Assessment Committee of Humanities at Radboud University.

Protection of participants' privacy

Protecting the privacy of individuals participating in a scientific experiment is crucial to preventing the misuse of the data they provide. All data collected were stored in

a separate file; each participant was pseudonymized using a key consisting of two letters and five digits to refer to participants' personal data in the files and datasets containing their responses in the intelligibility assessment. The file holding the personal data and the key linking to participant names was registered in accordance with the university's safety regulations for sensitive and critical data and is kept for a minimum of 10 years.

Data storage in the context of scientific integrity

During the listening experiment for speech intelligibility assessment, both personal data and judgments made by the participants were systematically collected using ROLEG and subsequently transferred to the storage facility within the Science Faculty at Radboud University, in accordance with the institute's policy. The storage location meets the legal and ethical requirements. The IT security and safety protocols guarantee the safe and secure storage. Furthermore, the audio files to be assessed in the listening experiments were selected from the existing JASMIN-CGN corpus. The chosen audio recordings were also saved in the exact location where the collected data was stored.

Selected speech recordings from the existing corpora: CGN, JASMIN-CGN, Librispeech, and L2-ARCTIC, were also stored at the secure storage facility of the Science Faculty. They were used for training acoustic models over a secure and encrypted network connection, utilizing OpenSSH technology.

Giving access to the data

This project is funded by the Chinese Scholarship Council (CSC) with Ref. No.201807110017. Although no specific requirements for sharing data were provided, it is generally encouraged to share and reuse research data. Therefore, research data were shared through the Radboud Data Repository (https://data.ru.nl). Except for the open-source corpora, the other collected data is available under restricted access, as I do not have permission from my participants to share the data publicly. In connection with this, we allow interested consortium partners access to the pseudonymized data available for reuse, provided that access has been agreed upon and the data is subsequently authorized by one of the collection owners. For more details, please visit https://data.ru.nl.

English Summary

The research presented in this doctoral dissertation addresses the multifaceted challenge of enhancing automatic assessment for second language speech. It specifically focuses on the critical dimensions of speech intelligibility and automatic pronunciation assessment. The study is driven by the practical limitations of conventional classroom instruction in which individualized and real-time feedback is often scarce. Consequently, this work investigates how computational methods can be optimized to deliver scalable and individualized educational experiences, serving as effective and reliable supplements to human tutors. The central thesis posits that robust assessment requires a shift from simple error detection to a more nuanced understanding of intelligibility and a deeper, linguistically informed modeling of mispronunciations. To this end, the dissertation is structured as a cumulative series of empirical studies that systematically explore the acoustic correlates of non-native speech, validate objective metrics for intelligibility, devise strategies for low-resource pluricentric languages, and propose novel articulatory-enhanced architectures for precise mispronunciation detection and diagnosis.

Chapter 2 initiates the investigation by examining the fundamental acoustic characteristics of second language (L2) speech. Prior to the development of complex assessment models, it was essential to identify the prominent acoustic-phonetic features that effectively differentiate non-native speech from native standards. Previous systems have frequently depended on generic spectral matching or obscure representations that fail to identify the underlying articulatory causes of learners' mistakes. This chapter aimed to pinpoint specific features indicative of deviations from native norms. Through a systematic classification experiment employing standardized acoustic-phonetic feature sets, the findings demonstrated that spectral characteristics (e.g., MFCCs and spectral flux), which determine the quality of individual phones, are undoubtedly significant. Nonetheless, they are not the exclusive factors influencing pronunciation quality. A crucial insight from this initial study was the substantial importance of temporal features. The analysis revealed that suprasegmental aspects, such as articulation rate, speech rate, and frequency, serve as potent discriminators between native and non-native speakers. This highlights the theoretical significance of temporal fluency and rhythm in second language acquisition. Therefore, this dissertation contends that precise computational modeling of L2 speech should extend beyond segmental features to incorporate temporal dynamics, thereby capturing the holistic nature of speech production.

Following this examination of acoustic-phonetic features, the research trajectory advances towards the construct of speech intelligibility. In L2 pronunciation

pedagogy, there has been a growing shift in instructional emphasis from strict accent reduction to prioritizing intelligibility, ensuring that the speaker's message is conveyed successfully. However, obtaining a reliable quantitative measure of intelligibility is challenging due to its subjective nature. Therefore, **Chapter 3** first evaluated the validity of distinct assessment paradigms to establish a robust ground truth for subsequent computational analysis. By conducting a rigorous comparison between subjective measures such as Visual Analogue Scale (VAS) ratings and objective measures derived from orthographic transcriptions (OTs), the research uncovered a vital distinction. The empirical data revealed only weak correlations between these two types of scores, suggesting they capture different facets of communication. Specifically, the VAS ratings were found to capture the listeners' intuitive perception of understanding, which is widely interpreted as reflecting comprehensibility, although such judgments may be influenced by listener attitude and cognitive load. In contrast, measures based on orthographic transcription, specifically word accuracy, provided a more direct and objective metric of functional intelligibility. Based on this validation, **Chapter 3** further investigated the automation of these metrics. By employing regression techniques to manage high-dimensional acoustic data, the study successfully demonstrated that the intelligibility scores could be accurately predicted using a selected subset of acoustic-phonetic features. This finding underscores the feasibility of automated systems in approximating human-level assessment reliability, offering a scalable alternative to the prohibitive time and labor demands of manual evaluation, thereby paving the way for scalable and automated intelligibility assessment.

Having established robust metrics and features, the dissertation addresses the technological bottleneck of data scarcity that is pervasive for non-dominant languages. While English enjoys extensive speech resources, pluricentric languages such as Dutch encounter a unique challenge regarding the scarcity of resources for non-dominant varieties like Flemish Dutch compared to their dominant counterparts such as Netherlandic Dutch. To bridge this gap, **Chapter 4** investigated whether the robustness of automatic speech recognition and pronunciation error detection systems for Flemish learners could be improved without incurring the prohibitive costs associated with collecting extensive new datasets. The research proposed and evaluated a strategy based on resource cumulation and transfer learning. Empirical results provided compelling evidence that cumulating data from different varieties significantly enhances the baseline performance of speech recognition systems. This approach effectively mitigates data scarcity by enabling the model to generalize across the pluricentric spectrum. Nonetheless, the study also uncovered a critical nuance concerning pronunciation error detection. While recognition performance

saw improvements, the precision of pronunciation error detection was compromised by systematic vowel differences between the varieties. The resulting mismatch caused the models to exhibit diminished discriminatory power, as the acoustic norms of the dominant variety did not correspond with the phonetic realizations of the non-dominant variety. This outcome indicates that, although data augmentation through variety cumulation is an effective strategy for conventional speech recognition tasks, mispronunciation detection systems depend more heavily on variety-specific modeling to align with the distinct phonetic norms of the target variety.

Chapter 5 aims to advance End-to-End (E2E) mispronunciation detection by integrating articulatory features to enhance feature representations. This approach aims to improve accuracy and facilitate detailed diagnostic feedback. Building upon the findings detailed in **Chapter 2,** which highlight the insufficiency of spectral features alone, this study addresses the limitation of traditional models that often fail to capture the physiological mechanisms necessary for practical learner guidance. In Section 5.1, the research develops two distinct output frameworks, defined as phoneme-based and articulatory label-based representations. The efficacy of these frameworks is examined using both customized Conformer-Transformer models and fine-tuned self-supervised XLSR models. The results indicate that while the incorporation of articulatory knowledge consistently improves overall detection performance, the explicit articulatory label-based output framework is particularly effective in maximizing diagnostic depth. Notably, the combination of this framework with XLSR embeddings achieved the highest performance in identifying subsegmental mispronunciations. Building upon these experimental findings, Section 5.2 conducts a comprehensive error analysis that elucidates the complexities inherent in the proposed approach. The findings reveal that, although models utilizing the articulatory label-based framework outperform those using phoneme-based configurations in diagnostic accuracy, they exhibit a slight decrease in detection precision. Additionally, these models show sensitivity to utterance length and inter-speaker variability. These observations suggest that, while the integration of articulatory knowledge represents a promising theoretical direction, future systems should incorporate more sophisticated mechanisms—such as hierarchical attention or cross-modal adaptive fusion—to ensure stable performance across diverse speaking conditions.

In conclusion, this dissertation presents a systematic examination of recent developments in automatic speech intelligibility and the assessment of L2 pronunciation. The research advances from identifying the fundamental acoustic signatures of non-native speech to validating objective methodologies for predicting

intelligibility scores, while also addressing the challenges posed by data scarcity in pluricentric languages. Moreover, it highlights the significance of incorporating linguistic knowledge, particularly articulatory features, into E2E architectures to improve the interpretability of models. By moving beyond conventional performance metrics to bridge the divide between signal processing and articulatory phonetics, this dissertation contributes to the development of intelligent systems capable of considering the underlying articulatory mechanisms of learners. Ultimately, these initiatives support the development of next-generation automated tutoring systems that deliver precise and diagnostically comprehensive feedback, transcending mere correctness scores to provide actionable insights, thus fostering more accessible and effective language acquisition.

Nederlandse Samenvatting

Het onderzoek dat in dit proefschrift wordt gepresenteerd, pakt de veelzijdige uitdaging aan om de automatische beoordeling van tweedetaalspraak (L2-spraak) te verbeteren. Het onderzoek richt zich daarbij specifiek op de cruciale dimensies van spraakverstaanbaarheid (speech intelligibility) en automatische uitspraakbeoordeling. De motivatie voor dit onderzoek wordt ingegeven door de praktische beperkingen van conventioneel klassikaal onderwijs, waarin geïndividualiseerde en real-time feedback vaak schaars is. De vraag is dan ook hoe computationele methoden kunnen worden geoptimaliseerd om bruikbare en geïndividualiseerde leerervaringen te leveren, die dienen als effectieve en betrouwbare aanvullingen op het werk van docenten. De centrale stelling luidt dat een bruikbare en stabiele computationele beoordeling een verschuiving vereist van een eenvoudige foutendetectie naar een genuanceerder begrip van verstaanbaarheid en naar een diepere, taalkundig onderbouwde modellering van uitspraakfouten. Om dit mogelijk te maken is het proefschrift opgebouwd als een cumulatieve reeks van empirische studies die systematisch de akoestische correlaten van niet-moedertaalspraak verkennen, objectieve maten voor verstaanbaarheid valideren, strategieën ontwikkelen voor pluricentrische talen met weinig bronnen (low-resource languages) en nieuwe architecturen voorstellen die verrijkt zijn met articulatorische informatie voor een nauwkeurige detectie en diagnose van uitspraakfouten.

Hoofdstuk 2 begint met onderzoek naar het bestuderen van de fundamentele akoestische kenmerken van tweedetaalspraak. Voorafgaand aan de ontwikkeling van complexe beoordelingsmodellen was het essentieel om de prominente akoestisch-fonetische kenmerken te identificeren die niet-moedertaalspraak effectief onderscheiden van moedertaalspraak. Eerdere systemen waren vaak afhankelijk van generieke spectrale matching of ondoorzichtige representaties die er niet in slagen de onderliggende articulatorische oorzaken van de fouten van leerders te identificeren. Dit hoofdstuk had tot doel specifieke kenmerken vast te stellen die indicatief zijn voor afwijkingen van moedertaalnormen. Door middel van een systematisch classificatie-experiment met gestandaardiseerde sets van akoestisch-fonetische kenmerken kon worden aangetoond dat spectrale kenmerken (bijv. MFCC's en spectrale flux), die de kwaliteit van individuele fonen bepalen, zonder meer van belang zijn. Desalniettemin zijn zij niet de enige factoren die de uitspraakkwaliteit beïnvloeden. Een cruciaal inzicht uit deze initiële studie was het substantiële belang van temporele kenmerken. De analyse onthulde dat suprasegmentele aspecten, zoals articulatiesnelheid, spreeksnelheid en frequentie, krachtige discriminatoren vormen tussen de spraak van moedertaalsprekers en die van niet-moedertaalsprekers. Deze uitkomst

benadrukt het belang van temporele vloeiendheid en ritme in tweedetaalverwerving. Daarom betoogt dit proefschrift dat nauwkeurige computationele modellering van L2-spraak verder moet gaan dan segmentele kenmerken en temporele dynamiek moet integreren, om zo de holistische aard van spraakproductie vast te leggen.

Na dit onderzoek naar akoestisch-fonetische kenmerken richt het onderzoekstraject zich op het construct van spraakverstaanbaarheid. De L2-uitspraakdidactiek schuift in toenemende mate van een instructieve aanpak met de nadruk op accentreductie naar een prioritering van verstaanbaarheid, waardoor beter wordt gewaarborgd dat de boodschap van de spreker succesvol wordt overgebracht. Het verkrijgen van betrouwbare kwantitatieve maatstaven voor verstaanbaarheid is echter uitdagend vanwege de subjectieve aard ervan. Daarom evalueert **Hoofdstuk 3** eerst de validiteit van verschillende beoordelingsparadigma's om een robuuste 'ground truth' vast te stellen voor daaropvolgende computationele analyses. Door een rigoureuze vergelijking uit te voeren tussen subjectieve beoordelingen met gebruikmaking van de Visual Analogue Scale (VAS) en objectieve maten afgeleid van orthografische transcripties (OT's) bracht het onderzoek een vitaal onderscheid aan het licht. De empirische data toonden slechts zwakke correlaties tussen deze twee typen scores, wat suggereert dat ze verschillende facetten van communicatie vastleggen. Specifiek bleken de VAS-beoordelingen de intuïtieve perceptie van begrip door de luisteraar vast te leggen, wat breed wordt geïnterpreteerd als een weerspiegeling van begrijpelijkheid (comprehensibility), hoewel dergelijke oordelen beïnvloed kunnen worden door de houding en cognitieve belasting van de luisteraar. Daarentegen boden maten gebaseerd op orthografische transcriptie, specifiek woordnauwkeurigheid (word accuracy), een directere en objectievere metriek voor functionele verstaanbaarheid. Op basis van deze positieve validatie onderzocht **Hoofdstuk 3** verder de automatisering van deze maten. Door regressietechnieken toe te passen op de omvangrijke verzameling van akoestische data kon succesvol worden aangetoond dat de verstaanbaarheidsscores nauwkeurig voorspeld konden worden met behulp van een subset van akoestisch-fonetische kenmerken. Deze bevinding onderstreept de bruikbaarheid van geautomatiseerde systemen om betrouwbaar de menselijke beoordeling te benaderen. Deze benadering biedt een bruikbaar alternatief voor de arbeidsintensieve en lastige inspanningen van handmatige evaluatie. Daarmee wordt de weg vrijgemaakt voor geautomatiseerde verstaanbaarheidsbeoordeling.

Na het onderzoek naar de rol van belangrijke spraakkenmerken en mogelijk automatische maten richt het proefschrift zich vervolgens op de technologische bottleneck van dataschaarste die alom aanwezig is bij het onderzoek naar niet-dominante talen. Terwijl het Engels beschikt over uitgebreide spraakbronnen,

geldt dat in beduidend mindere voor minder dominante talen. Een bijdrage aan de oplossing van dit probleem kan worden geput uit het bestaan van spraakbronnen in pluricentrische taalsituaties, zoals voor het Nederlands. Het Nederlands van Nederland heeft omvangrijkere spraakbronnen dan het Nederlands in Vlaanderen. Om deze kloof te overbruggen onderzoekt **Hoofdstuk 4** of de automatische spraakherkenning en detectie van uitspraakfouten van Vlaamse leerders verbeterd kan worden op basis van spraakbronnen van leerders uit Nederland. Het onderzoek stelt een strategie voor en evalueert deze, gebaseerd op de cumulatie van bronnen en transfer learning. Empirische resultaten leveren overtuigend bewijs dat het cumuleren van data van verschillende variëteiten de baseline-prestaties van spraakherkenningssystemen aanzienlijk verbetert. Deze aanpak vermindert effectief de gevolgen van dataschaarste door het model in staat te stellen te generaliseren over het pluricentrische spectrum. Desalniettemin brengt de studie ook een kritische nuance aan het licht betreffende de detectie van uitspraakfouten. Hoewel de herkenningsprestaties verbeterden, werd de precisie van de detectie van uitspraakfouten gecompromitteerd door systematische klinkerverschillen tussen de variëteiten. De resulterende mismatch zorgde ervoor dat de modellen een verminderd discriminerend vermogen vertoonden, aangezien de akoestische normen van de dominante variëteit niet overeenkwamen met de fonetische realisaties van de niet-dominante variëteit. Deze uitkomst laat zien dat, hoewel data-augmentatie door middel van variëteitscumulatie een effectieve strategie is voor conventionele spraakherkenningstaken, systemen voor de detectie van uitspraakfouten sterk afhankelijk zijn van variëteitspecifieke modellering om aan te sluiten bij de distincte fonetische kenmerken van de doelvariëteit.

Hoofdstuk 5 heeft tot doel End-to-End (E2E) detectie van uitspraakfouten te verbeteren door articulatorische kenmerken te integreren om daarmee kenmerk-representaties van klanken te verrijken. Deze benadering beoogt de nauwkeurigheid te verbeteren en gedetailleerde diagnostische feedback te faciliteren. Voortbouwend op de bevindingen in **Hoofdstuk 2**, die de ontoereikendheid van louter spectrale kenmerken benadrukken, richt deze studie zich op de beperkingen van traditionele modellen die er vaak niet in slagen de fysiologische mechanismen vast te leggen die nodig zijn voor de praktische begeleiding van de leerder. Het onderzoek ontwikkelt twee verschillende output-kaders, het ene gebaseerd op foneem representaties, het andere op articulatorische labels. De effectiviteit van deze kaders wordt onderzocht met behulp van zowel aangepaste Conformer-Transformer-modellen als 'fine-tuned' zelf-superviserende XLSR-modellen. De resultaten geven aan dat hoewel de integratie van articulatorische kennis de algehele detectieprestaties consistent verbetert, het expliciete kader gebaseerd op articulatorische labels bijzonder effectief

is in het maximaliseren van de diagnostische diepgang. Met name de combinatie van dit kader met XLSR-embeddings behaalt de hoogste prestaties in het identificeren van subsegmentele uitspraakfouten. Voortbouwend op deze experimentele bevindingen wordt een uitgebreide foutenanalyse uitgevoerd die de complicaties inherent aan de voorgestelde aanpak verheldert. De bevindingen laten zien dat modellen die het kader gebaseerd op articulatorische labels gebruiken weliswaar beter presteren in diagnostische nauwkeurigheid dan die met foneemgebaseerde configuraties, maar dat zij ook een lichte afname in detectieprecisie vertonen. Bovendien blijken deze modellen gevoelig voor uitingslengte en variabiliteit tussen sprekers. Deze observaties suggereren dat, hoewel de integratie van articulatorische kennis een veelbelovende theoretische richting vormt, toekomstige systemen geavanceerdere mechanismen zouden moeten incorporeren - zoals hiërarchische aandacht (hierarchical attention) of cross-modale adaptieve fusie - om stabiele prestaties in diverse spraakcondities te waarborgen.

Zoals uit de samenvatting van de voorgaande hoofdstukken blijkt, omvat dit proefschrift een systematisch onderzoek op grond van recente ontwikkelingen naar de automatische beoordeling van spraakverstaanbaarheid en L2-uitspraak. Het onderzoek reikt van het identificeren van de fundamentele akoestische kenmerken van niet-moedertaalspraak naar het valideren van objectieve methodologieën voor het voorspellen van verstaanbaarheidsscores, terwijl ook de uitdaging van dataschaarste in pluricentrische talen wordt aangepakt. Bovendien benadrukken de bevindingen het belang van het incorporeren van taalkundige kennis, met name articulatorische kenmerken, in E2E-architecturen om de interpreteerbaarheid van modellen te verbeteren. Door verder te gaan dan conventionele prestatiematen om de kloof tussen signaalverwerking en articulatorische fonetiek te overbruggen, draagt dit proefschrift bij aan de ontwikkeling van intelligente systemen die in staat zijn rekening te houden met de onderliggende articulatorische spraakproductieprocessen in tweedetaalleerders. Uiteindelijk ondersteunen deze onderzoeksinitiatieven en -uitkomsten de ontwikkeling van de volgende generatie geautomatiseerde tutorsystemen die precieze en diagnostische feedback zullen gaan leveren. Deze systemen overstijgen het karakter van louter correctheidsscores en zullen aan de taalleerders direct bruikbare en handelingsgerichte inzichten bieden wat zal leiden tot effectievere en toegankelijkere processen van tweedetaalverwerving.

Acknowledgements

It is deep winter in Beijing. As I type these lines, the chill outside my window seems to strangely overlap with the biting late autumn wind I felt when I first arrived in the Netherlands years ago. Dragging my suitcase, I stepped onto the land of Nijmegen for the first time in that cool season. The wind was cold, but my heart was burning with anticipation for the future. I never imagined that this journey, starting from the banks of the Waal River, would wind its way through such a long passage of time, finally reaching its conclusion here. This journey has spanned the Eurasian continent, traversed a global pandemic that changed the world deeply, and witnessed my transformation from a green youth into a man in his thirties. This doctoral thesis is not merely an academic summary of years of research in Computer-Assisted Language Learning, it stands more like a heavy milestone at the tail end of my youth, freezing in time those countless days and nights of oscillation between self-doubt and self-reinvention, and recording the names of those who were willing to hold me up when I was at my lowest.

Looking back on this long academic journey, I must first offer my highest respect and deepest gratitude to my supervisors: Helmer Strik, Roeland van Hout, and Catia Cucchiarini. As a team, you possess a perfect chemistry, combining a visionary macroscopic perspective, meticulous technical oversight, and humanistic guidance in academic writing. It is your tolerance, patience, and professionalism that built a pure academic sanctuary for me in this uncertain world.

I owe a special debt of gratitude to Helmer. For all these years, you have been not only my daily supervisor but the guiding light of my academic career, providing comprehensive oversight for all my research work. To this day, I still clearly remember the moment I received your first email reply and the subsequent phone interview. At that time, I was a master's student facing graduation, standing at a crossroads between finding a job and pursuing further academic research, filled with confusion and anxiety about the future. I did not even dare to believe I had the capability to pursue a PhD. It was you, on the other end of the phone, who affirmed me and extended the olive branch by accepting my application. This grace of recognition was the new starting point of my academic career and gave me my initial confidence. In the research work that followed, I knew I was not naturally rigorous or meticulous, often making careless mistakes in code or experimental design. Yet, you never scolded me for this. Instead, you always pointed out my errors with extreme patience and guided me step by step to find the right direction. Your rigorous attitude towards science and your inclusive embrace of your students is the most valuable lesson I learned during my PhD.

I am equally grateful to Roeland for clearing the fog for me at critical moments. You are a patient and professional mentor, and I am particularly impressed by your profound expertise in Statistical Linguistics and statistical analysis. Whenever I felt helpless in the face of complex experimental data, or when statistical results deviated from expectations, you always provided the most professional guidance and suggestions, helping me extract scientific conclusions from the chaotic data. Your macroscopic academic vision taught me to jump out of technical details and consider the deeper linguistic significance behind the research.

Catia, you played a crucial role in my doctoral research. In the experimental design phase, especially regarding the selection and processing of corpus resources, your professional advice laid a solid foundation for the subsequent model training. At the same time, I must specifically thank you for your tremendous help with thesis writing and academic expression. As a non-native English speaker, although I strove for accuracy, I often struggled to reach the level of a native speaker in the nuances of academic English and the flow of logical argumentation. With great patience and rigor, you helped me polish every manuscript. From the precision of word choice to the logical chain of arguments, your constructive feedback greatly improved the readability and professionalism of my thesis, helping me transform rough ideas into rigorous scientific language.

Beyond academic guidance, I also want to thank the unforgettable twists and turns of this journey. In late 2019, I returned to China to visit my family in Wuhan, thinking it would be a short vacation, only to encounter the lockdown of Wuhan in January 2020. In that instant, I was trapped in the eye of a historic storm. The lockdown, lasting nearly half a year, not only physically cut off my path back to the Netherlands but also psychologically devastated my research rhythm. I was at a stage where I most needed to produce results, yet outside the window were screaming ambulances and ever-increasing case numbers, while inside was a stagnant project and a panic over having zero output. Anxiety, insomnia, and even the thought of giving up struck me countless times during the late nights of that period. Thanks to the remote encouragement from my supervisors across the ocean and the companionship of my family, I was able to regain a sense of inner order amidst the chaos. I must also extend my heartful gratitude to Prof. dr. Mirjam Ernestus, the Director of the Centre for Language Studies (CLS) at that time. Even after the lockdown in Wuhan was lifted, returning to the Netherlands remained a daunting challenge. Upon learning of my predicament, Mirjam stepped in with incredible generosity and support, making it possible for me to return to Nijmegen. Her help cleared the final obstacle on my way back, allowing me to return to Nijmegen and immerse myself in my research as

soon as possible. The work completed during that time, though perhaps imperfect in hindsight, was a glimmer of light I earned for myself in a desperate situation, giving me the courage to carry on.

I would like to thank the Centre for Language and Speech Technology (CLST) for providing an open, friendly, and supportive environment that allowed me to complete my doctoral research. I am a relatively introverted person, especially in a new and strange environment, but everyone's enthusiasm and selfless sharing helped me overcome the initial unfamiliarity and broadened my cognitive boundaries both in life and in academic expertise. I am grateful to my officemate, Yu, for the days we spent discussing academic issues, life problems, and news in the office, which feel like yesterday despite the long period of working from home we experienced later. I thank Wei for your warmth in helping me actively integrate into the CLST family, and for inviting me to your 1 o'clock lunches with other colleagues when I was still apprehensive and overwhelmed. Your Mid-Autumn Festival mooncakes and hotpot gatherings meant a lot to me, and I also miss your two cute cats. Thank you, Wenwei, my junior alumnus from Beijing Language and Culture University, with whom I shared many common topics. Thank you, Xiaoru, for enthusiastically introducing me to your friends and helping me with many aspects of life. I still remember the pancake gathering and that evening around 10 P.M. in the small park in the city center, where we watched the sunset and discussed life. Thank you, Louis, for answering my questions with patience and enthusiasm, especially regarding the engineering practice in my experiments. Thank you, Ferdy and Daniëlle, for selflessly sharing the experimental data you collected to help me complete my doctoral work. Thank you, Aditya, for sharing your research insights and helping me resolve confusion regarding experimental details. Thank you, Henk, for your care and help throughout my PhD; knowing that "Henk knows everything" gave me great comfort whenever I was unsure about something. I also thank the members of the ICT development team at CLST, including Micha, Tim, and Wessel, for your essential assistance in carrying out my work. There are too many others to list, but I thank you all for the help you gave me during my PhD.

The years I lived in Nijmegen were a fantasy journey for me. There, I learned to cook, celebrated various Chinese and Dutch festivals, watched a football match in a stadium for the first time in my life, and traveled alone to a strange country for the first time. At this moment, images of Nijmegen surface in my mind, including the sunset by the Waal River, the lush woods on campus, and those sincere friends such as Jing, Changqing, Chen, Song, Jiangkun, Jieqiong, Zhaobao, Jiabin, Xin, Long, Lei, Xuan, Jinbiao, and so many others. I especially want to thank Jing and Changqing

for frequently inviting me to their homes for dinners and outings. Your warmth comforted my homesick heart in a foreign land, making those days abroad less monotonous and dull. Those years are the most colorful page in my memory where I saw different sceneries and met many wonderful people.

On this journey, I want to give special thanks to my wife, Jingping(静萍). From our first meeting in Beijing to our long-distance relationship while I studied in the Netherlands, and finally to our reunion in Beijing, our relationship has elevated from a couple to husband and wife. You have witnessed and borne the pressure brought by the delay of my studies. Thank you for being my emotional outlet countless times when I was down because of poor thesis progress, and thank you for taking care of this student who often fell into self-doubt and emotional swings, even while you were busy with your own work in Beijing. Thank you for your tolerance, understanding, and companionship all the way. Although this degree has arrived late, it belongs to both of us.

我要感谢我在武汉的父母，你们永远是我最无私的避风港。感谢你们给予我生命，从小到大，我好像从来都不是一个特别听话和懂事的孩子，但你们却给了我无尽的理解和包容，随着年龄的增长，我越来越能体会到你们的艰辛不易。感谢你们给予我支持，求学之路并非坦途，虽然经常让你们失望，但你们总是竭尽全力为我创造最好的环境，即便我说要远赴万里之外继续学业，你们心中虽有牵挂，却依然无条件地支持我，鼓励我。我依稀记得上小学时你们给我买的蓝色书包上绣着 “望子成龙” 这几个字，我总因没能成长为让你们骄傲的人而懊恼和心怀愧疚，但我一直告诉自己要努力，不要辜负你们的期待。2020 年封锁期间的朝夕相处我才注意到你们鬓角的白发，也才懂得了亲情的分量。

Finally, I want to say a few words to myself. I have been a person full of contradictions since I was young. I am sensitive and somewhat inferior, often sinking into the quagmire of self-doubt, feeling that I am not smart enough or doing well enough. But at the same time, I retain a bit of stubbornness in my heart, trying to prove myself amidst repeated negations. This ceaseless oscillation between self-denial and self-encouragement has run through my entire doctoral career. Frankly speaking, looking at this finally completed doctoral thesis before me, my mood is complex. It is not as perfect as I envisioned when I set out, nor has it made me an academic rising star that everyone, and myself, can be proud of. To reach this finish line, I spent too much time, paid too high a price, and arrived carrying a body full of exhaustion.

Facing these years of youth, facing the beautiful people and scenery I met in Nijmegen, and facing this ending that is not particularly outstanding, I would like to quote a lyric from a song I love very much as a final footnote: Though there is longing, and a trace of regret, all that remains is to say goodbye.

What I miss are these past years, especially that pure time in Nijmegen, the people I met, the scenery I saw, and the touched feelings I harvested. What I regret is the feeling that although it took so long to get here, the answer sheet I have handed in does not seem so excellent, and I have not become someone that everyone can be proud of, at least, that is what I think.

Nonetheless, this long farewell is finally completed.

Goodbye, Nijmegen.

Hello, future.

Curriculum Vitae

Xing Wei was born in Wuhan, Hubei province, China, in 1989. He obtained his Bachelor's degree in Computer Science and Technology from Huazhong Agricultural University Chutian College in 2012. Following his undergraduate studies, he decided to further specialize in his field. In 2015, he enrolled in the Master's program in Software Engineering at Beijing Language and Culture University. During this period, he developed a strong research interest in Computer-Assisted Language learning, exploring how engineering solutions can be applied to language acquisition. He received his Master of Engineering degree in 2018. In late 2018, supported by a scholarship from the China Scholarship Council (CSC), he moved to the Netherlands to pursue his PhD at Radboud University in Nijmegen. He conducted his doctoral research at the Centre for Language and Speech Technology, where he was a member of the Graduate School of Humanities. His research was supervised by Prof. Roeland van Hout, Dr. Helmer Strik, and Dr. Catia Cucchiarini. The results of his doctoral work are presented in this thesis.

List of publications

PhD related

Wei, X., Cucchiarini, C., van Hout, R., & Strik, H. (2025a). Articulatory-Enhanced Mispronunciation Detection and Diagnosis Models: A Multi-dimensional Error Analysis. *10th Workshop on Speech and Language Technology in Education (SLaTE)*, 31–35. https://d oi.org/10.21437/slate.2025-7

Wei, X., Dong, W., Cucchiarini, C., van Hout, R., & Strik, H. (2025b). Leveraging Articulatory Information to Enhance End-to-End Models for L2 Mispronunciation Detection and Diagnosis. *10th Workshop on Speech and Language Technology in Education (SLaTE)*, 76–80. https://doi.org/10.21437/slate.2025-16

Wei, X., Cucchiarini, C., van Hout, R., & Strik, H. (2023a). Measuring Intelligibility in Non-native Speech: The Usability of Automatically Extracted Acoustic-Phonetic Features. *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 121–125. https://doi.org/10.2 1437/slate.2023-23

Wei, X., Cucchiarini, C., van Hout, R., & Strik, H. (2023b). Distinctive Features for Classifying Spoken Native Versus Non-Native Speech. *9th Workshop on Speech and Language Technol-ogy in Education (SLaTE)*, 51–55. https://doi.org/10.21437/slate.2023-11

Wei, X., Hout, R. van, Catia Cucchiarini, Reuvekamp, D., & Helmer Strik. (2023c). Assessing Intelligibility in Non-native Speech: Comparing Measures Obtained at Different Levels. *Proc. Interspeech 2023*, 964–968. https://doi.org/10.21437/interspeech.2023-1635

Wei, X., Cucchiarini, C., van Hout, R., & Strik, H. (2022). Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: cumulating speech resources in a pluricentric language. *Speech Communication, 144*, 1–9. https://doi.org/10.1016/j.specom. 2022.08.004

Others

Xie, Y., Wei, X., & Wang, W. (2019). Manual and Semi-manual Comparison in the Annotation of CAPT Speech Corpus. *Journal of Physics: Conference Series,*1237, 032013. https://doi.org/ 10.1088/1742-6596/1237/3/032013

Wei, X., Chen, J., Wang, W., Xie, Y., & Zhang, J. (2017). A study of automatic annotation of PETs with articulatory features. *In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1608–1612. https://doi.org/10.1 109/apsipa.2017.8282281

9 789465 152967

Radboud Universiteit