The JALT CALL Journal
ISSN 1832-4215, 21(3), 102732 (2025)
https://doi.org/10.29140/jaltcall.v21n3.102732

Castledown

Systematic review of feature-based approaches to mispronunciation detection

Lakshani Nissanka
Informatics Institute of Technology, Sri Lanka
lakshani.20210570@iit.ac.lk

Banuka Athuraliya
Informatics Institute of Technology, Sri Lanka
banu.a@iit.ac.lk

K S Priyanayana
University of Moratuwa, Sri Lanka
sahan.p@iit.ac.lk

Accurate pronunciation is essential for successful communication in a second language (L2) as it significantly influences communicative effectiveness and perceived fluency. Mispronunciations frequently arise due to the influence of the learner's first language (L1), posing barriers to effective spoken communication. Therefore, pronunciation error detection (PED) has emerged as a critical research area within the domains of Computer-Assisted Language Learning (CALL) and Computer-Assisted Pronunciation Training (CAPT). Although numerous PED systems have been developed over recent decades, existing survey papers have mainly emphasized comparisons of modeling methodologies or learning paradigms, often neglecting the critical role of feature representation. To address this research gap, this survey introduces a novel, feature-based taxonomy for categorizing PED methodologies into four primary groups: Acoustic-based, Acoustic-Phonetic, Linguistic-based, and Hybrid approaches. Each category is systematically reviewed, summarizing over two decades of research work with respect to feature extraction techniques, modeling approaches, evaluation metrics, and the nature and quality of instructional feedback provided to learners. A detailed comparative analysis highlights significant trade-offs among these categories in terms of detection accuracy, interpretability, resource demands, and applicability in real-time or low-resource contexts. Furthermore, this survey discusses recent and emerging trends in PED research, including self-supervised learning frameworks, multimodal feature fusion, and integrating phonological knowledge with modern deep learning architectures. By synthesizing existing

knowledge and identifying gaps in current methodologies, this paper aims to provide clear insights and directions for future advancements in PED systems.

Keywords: Pronunciation error detection techniques, computer-assisted language learning, computer-assisted pronunciation training, feature-based taxonomy

Introduction

The ability to produce intelligible pronunciation is fundamental to successful second language (L2) acquisition, facilitating both formal and informal communicative interactions by ensuring that speech is smooth, comprehensible, and listener-friendly rather than merely rapid. Consequently, pronunciation errors can substantially hinder comprehensibility, even when a speaker has advanced grammatical and lexical proficiency, because they typically stem from the phonological structures of a learner's first language (L1).

Advances in speech technology and the growing accessibility of Computer-Assisted Language Learning (CALL) and Computer-Assisted Pronunciation Training (CAPT) platforms have fueled significant research interest in automatic pronunciation error detection (PED) (Stockwell, 2012). These systems aim to identify and localize deviations between learner speech and native norms, thereby enabling timely, targeted feedback.

CALL-based studies of pronunciation training indicate that technology-mediated instruction can enhance intelligibility, while also high-lighting challenges in feedback design and the influence of feature representation on pedagogical effectiveness (Metruk, 2024; Strik, Cucchiarini, & Truong, 2009). Nevertheless, PED remains inherently challenging due to the diversity of L2 learner profiles, the subjectivity of "acceptable" pronunciation, and the scarcity of annotated data, particularly for under-resourced languages.

Over the past two decades, PED systems have evolved from traditional forced-alignment and confidence-scoring approaches to sophisticated neural architectures with end-to-end modeling. While existing surveys have explored modeling techniques, error-diagnosis strategies, and various feedback mechanisms, there remains a notable lack of systematic categorization based on the types of features utilized in PED approaches. Such an emphasis is crucial, as feature selection directly affects model performance, feedback interpretability, and adaptability to diverse learning contexts.

This paper addresses this gap by introducing a feature-centric taxonomy of PED methods. We categorize systems into four primary groups: (1) acoustic-based, (2) acoustic-phonetic, (3) linguistic-based, and (4) hybrid approaches. By reviewing and comparing methods within these categories, we offer insights into the respective strengths, limitations, and trade-offs of each feature type. We also discuss overarching challenges such as real-time deployment, explainability, and data efficiency.

The remainder of this paper is organized as follows: Background provides foundational concepts in L2 phonology and typical pronunciation errors, and

Methodology discusses our approach to selecting and categorizing existing works. We then present four separate sections devoted to Acoustic-Based PED, Acoustic-Phonetic PED, Linguistic-Based PED, and Hybrid PED. Afterward, a Comparative Analysis consolidates key findings across these feature categories, and a Discussion and Future Directions addresses overarching issues and emerging trends. Finally, we offer a Conclusion that summarizes key points and suggests directions for future research.

Background

Pronunciation is a vital aspect of L2 proficiency, influencing intelligibility, speaker confidence, and listener perceptions (Derwing & Munro, 2005). Unlike errors in grammar or vocabulary, which may still allow for approximate communication, pronunciation errors can significantly impede understanding. For this reason, "acceptable pronunciation" is generally defined not in terms of native-like accuracy, but by a learner's ability to produce speech that remains intelligible and comprehensible to listeners, even when accented. These errors commonly fall into two categories: segmental and suprasegmental.

Segmental errors involve inaccuracies in producing individual consonants and vowels, often due to L1 transfer. For instance, L2 learners from language backgrounds that lack the English dental fricative /θ/ may substitute it with /t/ or /s/, yielding "tink" or "sink" instead of "think," thus reducing intelligibility (Ehrlich & Avery, 2013) By contrast, suprasegmental errors concern stress, intonation, pitch variation, and rhythmic patterns. Misplaced stress or unnatural intonation can obscure intended meaning, even if individual sounds are correctly articulated (Wennerstrom, 2001; Yağız et al., 2024).

Numerous psycholinguistic factors contribute to the persistence of these errors. Phonological interference from the L1 is particularly salient, as learners often perceive and produce unfamiliar L2 sounds according to their existing phonemic categories (Aoyama et al., 2004) The age at which L2 learning begins also matters younger learners generally show greater neuroplasticity, facilitating more accurate pronunciation (Flege, 1995; Saito, 2022). Additionally, many L2 classrooms struggle to offer sufficient practice opportunities or individualized corrective feedback, exacerbating a "perception-production gap" where learners can recognize a phoneme in listening tasks but still fail to produce it accurately in speech (Munro & Derwing, 1995; Trofimovich & Baker, 2006).

Against this backdrop, CAPT systems increasingly rely on automatic detection of pronunciation deviations to provide targeted, real-time feedback. A key design factor in these systems is the feature representation extracted from learner speech. The choice of features, ranging from raw acoustic signals to linguistic or articulatory descriptors, shapes how effectively the system identifies, localizes, and explains errors. In the next section, we categorize and survey PED approaches according to their primary feature types, highlighting how these choices influence performance, interpretability, and scalability.

Methodology

This survey adopts a feature-centric framework to categorize and analyze PED systems. Rather than organizing prior work by modeling paradigm or application context, we classify systems based on the types of features they utilize acoustic, acoustic-phonetic, linguistic, and hybrid. This taxonomy enables a more principled understanding of how input representations influence system behavior, performance, and pedagogical relevance.

Literature selection

Relevant literature was gathered through a comprehensive search of academic databases, including Google Scholar, Scopus, and IEEE Xplore. The search covered studies published between 1995 and 2024. Search terms included combinations of "pronunciation error detection," "mispronunciation detection," and "computer-assisted pronunciation training" (CAPT). Reference chaining was also employed to ensure the inclusion of foundational and frequently cited studies.

Inclusion criteria

To be included in the survey, studies had to meet the following criteria:

- Focus explicitly on automatic pronunciation error detection or scoring in the context of L2 learning.
- Provide sufficient detail on the types of features extracted from learner speech and how those features are used in modeling or feedback generation.
- Be peer-reviewed journal or conference publications, preferably within the fields of speech processing, natural language processing, language education, or human-computer interaction.
- Address segmental or suprasegmental pronunciation errors.

Review process

Each selected study was reviewed and categorized according to its dominant feature type. The following aspects were extracted from each paper to enable consistent comparison:

- The type of feature representation used in the system.
- The modeling techniques applied.
- The feedback mechanisms implemented by the system.
- The evaluation metrics used to assess performance.
- The deployment context for which the system was designed.

The review followed general principles of systematic survey methodology, with studies screened at the title, abstract, and full-text levels. Categorization decisions were discussed among the authors to reduce subjectivity and ensure alignment. This approach enabled a structured comparison of systems and

highlighted trade-offs among feature complexity, feedback effectiveness, and scalability. By foregrounding feature representation, the survey links technical system design to pedagogical outcomes in L2 learning.

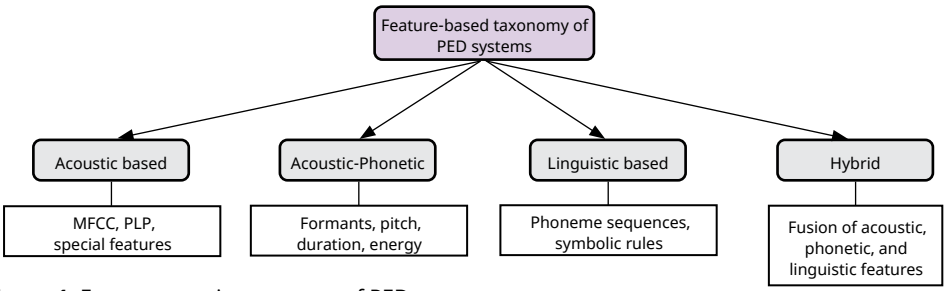

Figure 1. Feature-centric taxonomy of PED systems

Acoustic features-based PED

Acoustic-only approaches to PED are characterized by their exclusive reliance on low-level signal-derived features such as MFCCs, pitch, energy, and formant trajectories. These methods operate without the use of symbolic linguistic representations or phoneme-level alignments, focusing instead on the raw acoustic properties of learner speech to detect deviations from native-like pronunciation. In contrast to acoustic-phonetic or linguistically informed systems, acoustic-only models do not utilize phoneme posterior probabilities, forced alignment, or explicit phonetic labels. As a result, they offer practical advantages in scenarios where linguistic resources or annotation efforts are limited, including early-stage assessment tools, mobile learning applications, and language environments where expert phonetic annotations are unavailable.

One notable example of an acoustic-only PED system is presented by Chhabra et al. (2023), who employed MFCCs extracted from learner and native speaker speech to train a support vector machine (SVM) classifier for mispronunciation detection. Rather than aligning speech to phoneme sequences or generating phoneme-level labels, their system operated directly on frame-level acoustic features, classifying entire segments based on the similarity of their acoustic patterns. The SVM learned to distinguish correct from incorrect pronunciations without the need for linguistic supervision, demonstrating the ability of machine learning algorithms to extract useful patterns solely from acoustic data. Similarly, Shahin and Ahmed (2019) proposed an anomaly detection approach that builds acoustic models using only native speaker data, treating deviations in learner speech as outliers. Their method trained statistical models on MFCCs representing correctly pronounced speech and used measures such as reconstruction error and divergence-based distances to flag mispronunciations. This eliminates the requirement for labeled non-native data, making the system particularly suitable for resource-scarce environments and unsupervised learning contexts.

Building on the success of statistical models, Hu et al. (2015) applied deep learning to the task of pronunciation assessment using only acoustic inputs.

Their system trained deep neural networks (DNNs) using low-level features like MFCCs and perceptual linear prediction (PLP) coefficients to classify pronunciation quality. Although the broader framework of their work includes transfer learning and supervised adaptation, the DNN-based acoustic model itself operates on raw speech features without explicit phoneme labels. The model demonstrated superior performance compared to Gaussian mixture model (GMM) baselines, confirming that deep architectures can learn complex acoustic patterns directly from raw features. Importantly, this study highlights the effectiveness of data-driven learning from acoustic representations in scenarios where linguistic supervision is limited or unavailable.

In another important contribution, Bartelds et al. (2020) introduced a new pronunciation distance measure based purely on acoustic similarity. Their method compared MFCC-derived time-aligned sequences from learner speech and native references to quantify pronunciation differences without relying on phoneme-level segmentation or annotation. The distance metric was used to provide gradient feedback, indicating the degree to which a learner's pronunciation approximates a native-like form. This type of continuous scoring, based on acoustic distance rather than categorical correctness, offers nuanced insights into learner performance. Unlike binary classification systems, this approach supports formative assessment, helping learners monitor gradual improvement without requiring phonetic interpretation.

In addition to these approaches, Byun and van der Haar (2019) proposed a pronunciation detection system for foreign language learners that relies solely on MFCC-based acoustic features. Their method applied a Support Vector Machine (SVM) classifier to distinguish between correctly and incorrectly pronounced utterances without using phoneme alignment or symbolic linguistic representations. The system was trained on frame-level MFCC vectors extracted from learner speech and used supervised learning to classify pronunciation quality. Crucially, their work avoided the use of phoneme boundaries, forced alignment, or phoneme-level feedback, making it a clear example of an acoustic-only PED system. The study demonstrated that with careful feature selection and classifier design, acoustic-only models could effectively detect mispronunciations in a resource-efficient manner, even in the absence of detailed linguistic annotations. This reinforces the idea that low-level acoustic representations, when paired with statistical classifiers, can provide scalable and lightweight solutions for pronunciation assessment in computer-assisted language learning environments.

Further demonstrating the pedagogical applicability of acoustic-based modeling, C.H. Lee and Hsieh (2017) examined an ASR-driven pronunciation learning system integrated into an L2 classroom setting. Their study found that learners who practiced with automatic speech recognition feedback achieved significant gains in segmental accuracy compared with those receiving traditional instruction. Because the system evaluated pronunciation solely from acoustic recognition scores, it exemplifies how low-level signal features can effectively support formative assessment in CALL environments without relying on detailed phonetic labeling or linguistic modeling. This integration of

acoustic scoring within a pedagogical framework highlight how ASR-based systems can bridge the gap between automatic detection and meaningful learner improvement.

The effectiveness of acoustic-only approaches highlights the central role of feature representation in pronunciation error detection, demonstrating that even without linguistic supervision, carefully selected acoustic features such as MFCCs, pitch, and energy can capture meaningful deviations from native-like pronunciation. These methods are particularly valued for their scalability, simplicity, and minimal reliance on annotated data, making them well-suited for mobile learning, early-stage assessment, and low-resource environments. However, these benefits often come at the cost of limited interpretability and diagnostic power. Without phoneme-level alignment or symbolic representations, acoustic-only systems typically provide binary or coarse feedback, making it difficult to localize specific pronunciation errors or offer detailed correction guidance. Additionally, while some models require substantial training data or native reference utterances, others may struggle to generalize across diverse speakers or accents. Thus, while acoustic-only methods offer practical and resource-efficient solutions, their effectiveness in pedagogical contexts may be constrained by their inability to deliver fine-grained, interpretable feedback.

Table 1. Summary of acoustic-based PED approaches

Study	Features	Modeling technique	Strengths	Limitations
Chhabra et al. (2023)	Frame-level MFCCs	SVM classifier	Simple design, no phoneme alignment needed.	Minimal interpretability, binary feedback.
Shahin and Ahmed (2019)	MFCCs, energy	Unsupervised anomaly detection	Works without labeled non-native data, good for low-resource settings.	Hard to localize specific errors, and there is limited feedback.
Bartelds et al. (2020)	MFCC-derived distance measures	Time-aligned acoustic similarity	Continuous scoring and useful for incremental improvement.	Requires native reference utterances.
Byun and van der Haar (2019)	Frame-level MFCC vectors	SVM classification	Lightweight, minimal linguistic resources.	Low interpretability, no specific phoneme-level diagnostics.
Hu et al. (2015)	MFCCs, PLPs	DNNs	Learns rich features; outperforms GMM baselines.	Requires substantial training data, no explicit alignment for error localization
C.H. Lee and Hsieh (2017)	Acoustic recognition scores (ASR-based features)	ASR integrated into a pronunciation learning system	Demonstrated significant learner gains in segmental accuracy, scalable for CALL applications without detailed phonetic annotation.	Limited interpretability of feedback; system performance depends on ASR accuracy and robustness to accented speech.

Acoustic-phonetic features-based PED

Acoustic-phonetic approaches in PED combine low-level acoustic analysis with phonetic knowledge to identify and diagnose mispronunciations. These systems go beyond simple signal processing by aligning acoustic features with expected phonological targets, most often at the phoneme level, allowing for detailed comparisons between learner speech and native speaker models. By interpreting acoustic features such as formants, MFCCs, energy, and pitch through a phonetic lens, these methods can provide fine-grained articulatory feedback and more interpretable diagnostic information than purely acoustic models. Within a feature-based taxonomy, acoustic-phonetic systems represent a critical middle ground, offering a balance between raw signal processing and symbolic linguistic representation.

An early CALL-oriented example of this integration is provided by Hincks (2003), who demonstrated how speech technologies could visualize learners' pitch, energy, and timing contours to deliver feedback on prosody and intonation. Conducted within a pronunciation training environment, this work illustrates how acoustic-phonetic analysis can translate directly into pedagogically useful feedback, linking measurable acoustic cues with communicatively relevant aspects of speech such as stress and rhythm. Hincks's findings established a foundation for subsequent research that sought to combine phonetic accuracy with real-time, interpretable feedback in CAPT.

One of the earliest acoustic-phonetic approaches in this domain was introduced by Eskenazi (1999), who applied Automatic Speech Recognition (ASR) techniques to detect pronunciation errors in non-native English learners. Her system used forced alignment to map learner speech onto native speaker phoneme models and extracted acoustic likelihood scores to evaluate deviations. Although the system operated on signal-derived features, the analysis was conducted at the phoneme level, which reflects a clear integration of acoustic and phonetic information. This early work laid the groundwork for phoneme-level scoring by showing that forced-alignment-based likelihoods could reveal subtle mispronunciations, though its accuracy depended heavily on the quality of the ASR system and the robustness of phoneme models.

Building on this foundation, Witt and Young (2000) introduced the now widely adopted Goodness of Pronunciation (GOP) score, a likelihood-based metric for evaluating phoneme-level pronunciation quality. The GOP score is calculated as the log-likelihood of the observed acoustic feature sequence Ogiven the canonical phonemep,subtracted by the average log-likelihood across a set Q competing phonemes:

$$GOP(p) = log\ P(O|p) - \frac{1}{N}\sum_{q \in Q} log\ P(O|q)$$

Where P(p)is the likelihood of the observed features for the target phoneme, and P(q) represents the likelihoods for N alternative phonemes in the set Q. A higher GOP score indicates a closer match between the learner's pronunciation and the native model for the expected phoneme. This method operates directly

on acoustic models, typically Hidden Markov Models (HMMs), but interprets the resulting scores through a phonemic lens, making it a prototypical example of acoustic-phonetic feature integration. The GOP framework demonstrated how carefully engineered features, aligned with phonetic expectations, could produce interpretable scalar scores that correlate strongly with expert human assessments, thereby laying the groundwork for many subsequent systems in pronunciation scoring.

To enhance modeling robustness, Qian et al. (2012) proposed a hybrid Deep Belief Network-Hidden Markov Model (DBN-HMM) framework for mispronunciation detection. This system incorporated acoustic features such as MFCCs and Perceptual Linear Predictions (PLPs) but enriched their representational power through unsupervised pretraining and deep architectures. The phoneme-level output remained central to the system, allowing it to map raw features onto articulatory targets while improving resilience to acoustic variability and speaker diversity. This shift from shallow statistical models to deep neural representations illustrated how improved feature encoding could lead to more accurate phonetic interpretation without sacrificing interpretability.

Similarly, Lee and Glass (2015) advanced this paradigm by proposing a system that identifies mispronunciations without requiring annotated non-native speech data. Their GMM-HMM-based approach analyzes pronunciation consistency and error patterns using Extended Recognition Networks (ERNs) and speaker adaptation. Importantly, while their model processes acoustic features such as MFCCs, it interprets them through phoneme-level templates, retaining the acoustic-phonetic nature of the system. The model's capacity to iteratively refine phoneme-level error candidates without non-native supervision demonstrated how leveraging acoustic-phonetic features could enhance scalability and adaptability across learning contexts.

In parallel to these foundational systems, several other studies have explored the use of discriminative acoustic-phonetic features to detect phoneme-specific mispronunciations. For example, Truong et al. (2004) focused on detecting common errors among Dutch L2 learners by designing features informed by phonetic theory, such as formant trajectories and rate of rise, targeted toward phonemes with known substitution patterns. Their results underscored how the selection of features aligned with phonological distinctions could significantly improve classification accuracy and diagnostic clarity.

An important comparative study by Strik et al. (2009) evaluated the effectiveness of different classification approaches for detecting L2 mispronunciations, particularly the substitution of the Dutch velar fricative /x/ with the velar plosive /k/. The authors tested four models: two based on acoustic–phonetic features with Linear Discriminant Analysis (LDA), one using MFCCs with LDA, and a confidence-based model employing the GOP score. Their results showed that the acoustic-phonetic classifier with LDA achieved the highest accuracy of up to 93%, significantly outperforming the GOP-based model. These findings highlight the advantage of incorporating phonetic knowledge into feature selection and classification, enabling more precise identification of specific articulatory errors that may not be detectable through generic confidence measures alone.

Recent work has also extended acoustic-phonetic approaches using neural architectures. Jenne and Vu (2019) introduced a neural-based approach for pronunciation error detection that incorporates both acoustic and phonetic information. Their system processes aligned learner speech using MFCC vectors and spectrogram images to predict phoneme-level labels, which are then compared against expected outputs to generate corrective feedback. The architecture demonstrates how combining different acoustic representations with phonetic interpretation can enhance error detection accuracy while supporting more targeted feedback. This approach exemplifies the acoustic-phonetic feature category, where signal-level data is interpreted through phonological categories for improved diagnostic value.

In the realm of end-to-end modeling, Yan and Chen (2021) explored a fully end-to-end neural model for mispronunciation detection and diagnosis that processes learners' speech directly from raw waveforms. Their model employed the SincNet module, which utilizes learnable bandpass filters to convert raw waveforms into suitable vector representations. This approach retained more acoustic phenomena compared to conventional handcrafted features, potentially aiding neural networks in discovering more customized representations. Experiments conducted on the L2-ARCTIC dataset demonstrated that the model achieved performance comparable to state-of-the-art models using standard handcrafted acoustic features, with notable improvements in phone error rate and diagnosis accuracy.

Do et al. (2024) proposed an acoustic feature mix-up strategy designed to improve pronunciation assessment performance under data scarcity and label imbalance. Their method introduces both linear and nonlinear mixup techniques applied to acoustic representations, specifically those based on GOP scores. In addition to these augmentation methods, they incorporate fine-grained phonetic error-rate features by comparing automatic speech recognition outputs with reference phoneme sequences, allowing the model to capture detailed mispronunciation patterns. While the mixup strategy primarily enhances the robustness of acoustic modeling, the integration of phoneme-level error analysis introduces a phonetic dimension to the system's feedback and scoring mechanisms. This work demonstrates how acoustic and phonetic information can be jointly leveraged to improve the accuracy and generalization of pronunciation error detection systems, particularly in resource-constrained learning environments.

Acoustic-phonetic approaches occupy a vital middle ground in pronunciation error detection by combining low-level acoustic features with phoneme-level phonetic knowledge, enabling more interpretable, diagnostic feedback than purely acoustic methods. These systems leverage forced alignment, likelihood-based scoring, and phoneme-level classifiers to map learner speech to native-like articulatory targets, offering fine-grained assessments that correlate well with human ratings. Their strengths lie in their ability to provide localized, phoneme-specific feedback and to incorporate phonetic theory into feature design, improving detection of substitution errors and articulatory deviations. However, these benefits often come with increased complexity,

including dependence on accurate ASR alignment, labeled phoneme data, and more computationally intensive models. Additionally, while many models offer strong phoneme-level insights, they may fall short in handling prosodic or suprasegmental features, and their effectiveness can be sensitive to speaker variation, accent coverage, and resource availability. Despite these challenges, the structured interpretability and diagnostic precision of acoustic-phonetic systems make them indispensable for learner-aware, pedagogically meaningful pronunciation assessment tools.

Table 2. Summary of acoustic-phonetic based PED approaches

Study	Features	Modeling technique	Strengths	Limitations
Hincks (2003)	Pitch, energy, and timing contours	Speech technology-based prosody visualization and feedback	Provided interpretable feedback on stress and intonation.	Limited to prosodic aspects; lacks segmental error detection.
Eskenazi (1999)	MFCCs + forced alignment	ASR-based forced alignment, likelihood scoring	Early phoneme-level scoring, interpretable feedback.	Dependent on ASR performance, limited accent coverage.
Witt and Young (2000)	Acoustic features → phoneme (GOP)	HMM-based alignment + Goodness of Pronunciation	Well-established baseline, correlates with human ratings.	Relies on robust HMMs may miss prosodic errors.
Qian et al. (2012)	MFCCs, PLPs + phoneme alignment	Deep Belief Network-HMM hybrid	More accurate than GMM-HMM and robust to variability.	Complex training and requires labeled phoneme data.
Lee and Glass (2015)	MFCCs with phoneme-level templates	GMM-HMM, ERNs	Iterative refinement without labeled non-native data.	Must rely on reliable alignment, limited prosodic handling.
Do et al. (2024)	GOP scores + phonetic error-rate features	Acoustic mixup + fine-grained phoneme-level analysis	Works with limited data, detailed phonetic feedback.	Complex augmentation strategy, sensitive to alignment errors.
Truong et al. (2004)	Symbolic phoneme-level representations	Phonetic theory Informed classification	Targets known problematic phones and improves detection of L2 errors.	Depends on expert rules, may not handle new/unknown errors.
Strik et al. (2009)	Phoneme sequences + GOP-based confidence	LDA with confidence measures	Combines acoustic & symbolic info, high accuracy for known errors.	Requires forced alignment, lacks prosodic/intonation coverage.
Jenne and Vu (2019)	MFCCs + spectrograms → phoneme labels	CNN-based spectrogram + MFCC integration	Fuses multiple acoustic views and enables phoneme-level diagnosis.	Requires aligned data and complex architecture.
Yan and Chen (2021)	Raw waveforms + learned SincNet filters	End-to-end SincNet + CNN	Custom feature learning and improved error diagnosis accuracy.	Requires substantial data and black-box interpretability.

Linguistic features-based PED

Linguistic-based approaches to PED operate primarily on symbolic representations of speech, such as phoneme sequences, syllable structures, and rule-based transcription systems. These methods emphasize the role of language-specific knowledge and phonological structure in identifying and explaining learner errors. Rather than focusing on low-level acoustic signals, linguistic-based systems derive features from phoneme-level outputs, often generated by an ASR system, and compare them to canonical transcriptions to detect deviations. By incorporating linguistic theories of segmental and suprasegmental contrast, phonological transfer, and common L1-L2 mismatches, these systems offer a more interpretable and pedagogically aligned framework for diagnosing pronunciation errors. Importantly, the strength of this category lies in its ability to abstract away from acoustic variability and instead highlight recurrent, rule-governed patterns of mispronunciation that are often overlooked by acoustic or statistical models.

One early contribution to this domain was made by Menzel et al. (2000), who developed a constraint-based diagnosis system that performs automatic detection and correction of non-native English pronunciations. The system relies on symbolic representations derived from phoneme recognition and applies a set of predefined linguistic rules to detect deviations from canonical transcriptions. Rather than using acoustic scores or probabilistic models, the approach interprets learner input through a symbolic lens, identifying insertions, deletions, and substitutions in the phoneme string. By focusing on linguistic structure and error patterns, the system provides learners with intelligible, rule-based feedback that is pedagogically aligned and free from the opacity of black-box acoustic models.

Building on the role of cross-linguistic influence in pronunciation learning, Helen Meng et al. (2007) proposed a framework to derive salient pronunciation errors in Cantonese-speaking learners of English through symbolic phoneme-level analysis. Their approach compares phoneme sequences generated by an ASR system with canonical English transcriptions, highlighting frequent segmental substitutions based on predicted L1-L2 phonological contrasts. These contrasts are embedded in an extended pronunciation lexicon, which encodes likely mispronunciations informed by phonological distance and transfer patterns. The analysis is conducted at the symbolic level by evaluating frequency-driven error patterns across multiple instances, allowing researchers to isolate and interpret systematic learner errors without resorting to acoustic feature modeling.

Further examining the role of linguistic structure in pronunciation learning, Neri et al. (2006) evaluated an ASR-based CAPT system designed to provide feedback on problematic Dutch phonemes for non-native speakers. While the system uses ASR to transcribe learner input, the feedback mechanism operates by comparing the output against target phoneme sequences and predefined linguistic criteria. The study demonstrates that segmental pronunciation errors can be effectively addressed using structured linguistic feedback, confirming that symbolic comparison at the phoneme level, rather than acoustic modeling

alone, can yield pedagogically meaningful interventions. Their findings suggest that phonological knowledge, when explicitly encoded into feedback systems, contributes significantly to pronunciation improvement among language learners.

Recent CALL-focused research reinforces the pedagogical importance of linguistic and symbolic feedback in pronunciation training. Yonesaka (2017) examined learner perceptions of peer pronunciation feedback using the online P-Check platform, showing that structured phoneme-level comparisons increased learners' awareness of pronunciation strengths and weaknesses. Their findings indicate that explicitly presenting linguistic representations, such as phoneme transcriptions or categorized error types enhances both learner engagement and understanding in computer-assisted pronunciation contexts. These results highlight that learners benefit not only from acoustic feedback but also from interpretable, rule-based cues that help them connect performance outcomes to specific linguistic targets.

In addition, Amrate & Tsai (2024) conducted a systematic review of computer-assisted pronunciation training studies and reported that systems incorporating rule-based or linguistically grounded feedback promote deeper learner reflection and more sustained improvement than those relying solely on acoustic scoring. Their analysis underscores a growing trend in CALL research toward integrating explicit linguistic information, such as phonological rules or cross-linguistic contrast patterns into pronunciation error detection and feedback mechanisms. Collectively, these studies demonstrate that linguistic-based PED systems align closely with CALL's instructional goals by providing feedback that is both pedagogically meaningful and cognitively interpretable for learners.

Overall, linguistic-based PED systems demonstrate that symbolic feature representations grounded in phonological theory can serve as a powerful tool for interpreting learner speech. These approaches excel at capturing systematic pronunciation patterns influenced by a learner's L1 and are well-suited for generating structured, intelligible feedback that aligns with instructional goals. While they may not match the granularity of acoustic-based methods in scoring subtle articulatory deviations, their strength lies in error interpretability, pedagogical clarity, and cross-linguistic insight. However, these systems often depend heavily on the accuracy of ASR-generated transcriptions, which can introduce noise and misinterpretation into symbolic analyses. Additionally, their limited capacity to capture prosodic or fine-grained articulatory features may reduce their sensitivity to more nuanced errors in learner speech. Within a feature-based taxonomy, linguistic approaches remind us that the quality of a system's output is often shaped not only by data or modeling technique but by how well the input features reflect the linguistic realities of second language acquisition.

Table 3. Summary of linguistic-based PED approaches

Study	Features	Modeling technique	Strengths	Limitations
Menzel et al. (2000)	Phoneme sequences, symbolic rules	Constraint-based diagnosis	Rule-based, interpretable, and identifies insertions/deletion.	Dependent on ASR accuracy, may overlook subtle acoustic cues.
Helen Meng et al. (2007)	ASR-derived phoneme sequences, extended pronunciation lexicon	Comparison with canonical forms	Models L1–L2 contrasts, systematic detection of substitutions.	Relies on robust ASR and limited suprasegmental analysis.
Neri et al. (2006)	ASR-generated phoneme sequences	Edit-distance scoring, symbolic error rules	Structured feedback at phoneme level, useful in classrooms.	Vulnerable to misrecognition, minimal prosodic handling.
Yonesaka (2017)	Phoneme-level symbolic feedback via P-Check	Online peer pronunciation feedback system	Increased learner awareness of pronunciation strengths and weaknesses.	Dependent on peer input quality, limited automation.

Hybrid feature-based PED

Hybrid approaches in PED have emerged as a dominant trend in recent research, where systems integrate acoustic, phonetic, and linguistic features to capitalize on the strengths of each modality. These approaches have demonstrated considerable improvements in both detection accuracy and diagnostic interpretability, particularly in handling diverse learner pronunciations and varied speech conditions. Most contemporary studies now adopt some form of multi-feature integration, whether through architectural fusion, joint optimization, or layered supervision, reflecting a broader shift away from isolated feature pipelines. Deep learning frameworks, especially end-to-end models, have further accelerated this trend by enabling the seamless incorporation of heterogeneous features while minimizing dependence on manually engineered preprocessing.

Early efforts toward this integration can be seen in Zhang et al. (2020), who proposed an integrated framework that brings together traditional acoustic modeling and classification components for pronunciation error detection. The system is composed of both offline and online processing stages. During the online phase, learner speech is passed through a feature extraction module followed by forced alignment and a series of confidence scoring techniques such as GOP, Log-Likelihood Ratio (LLR), and Log-Posterior Probability (LPP). These scores are then classified to determine whether a segment is correct or contains a mispronunciation. This architecture reflects a hybrid design as it incorporates both signal-level acoustic features and phoneme-level confidence measures while leveraging expert scoring databases and classification models trained offline. It highlights the evolution from simple forced-alignment

systems to more comprehensive pipelines that combine data-driven acoustic processing with linguistically informed evaluation metrics.

Building on this, Lo et al. (2020) introduced an end-to-end mispronunciation detection system that eliminates the need for phoneme-level forced alignment. Their architecture combines convolutional and recurrent layers with an attention mechanism to learn error patterns directly from raw waveforms. Although primarily signal-driven, the model implicitly captures phonetic information through the temporal patterns it encodes, making it a functional hybrid despite not relying on hand-engineered phonetic features.

To mitigate the challenges posed by limited labeled data from L2 learners, Lo et al. (2021) proposed two augmentation strategies that enhance the discrimination capability of end-to-end mispronunciation detection models. Their input augmentation strategy distills phonetic discrimination knowledge from a Deep Neural Network Hidden Markov (DNN-HMM) acoustic model, effectively enriching the system with phonetic insights derived from supervised acoustic modeling. Simultaneously, the label augmentation strategy captures linguistic and phonological patterns from training transcripts, embedding structured symbolic information into the model's learning process. While the architecture remains end-to-end, these augmentations implicitly incorporate acoustic, phonetic, and linguistic representations, yielding improved robustness against overfitting in low-resource settings.

In a similar way, Chilaka (2022) developed an end-to-end mispronunciation detection system tailored for L2 English learners, which utilizes acoustic features extracted via convolutional and recurrent layers alongside linguistic-level supervision through attention-based decoders. The model integrates a CTC loss to align speech segments with expected output sequences and employs an attention mechanism over either character or phoneme-level units, enabling the system to capture structured linguistic patterns from learner speech. By incorporating both signal-level representations and symbolic transcriptions, the model demonstrates improved performance in detecting segmental pronunciation errors such as substitutions, insertions, and deletions, offering a more interpretable framework than systems relying solely on acoustic modeling.

Recently, Anwar et al. (2022) proposed a non-autoregressive end-to-end model for pronunciation error detection that focuses on improving inference efficiency while maintaining accuracy. The architecture consists of two complementary components: a dictation model, which generates token sequences from speech using acoustic features, and a pronunciation model, which evaluates the correspondence between these outputs and annotated phoneme sequences. By separating generation and verification, the model reduces sequential dependencies during decoding, allowing for parallelized inference. Importantly, the approach combines low-level acoustic signal processing with symbolic linguistic supervision, enabling the detection of pronunciation errors through both waveform-derived representations and structured phonetic comparisons. This design supports fast, interpretable error detection, suitable for real-time applications in second-language learning contexts.

Shahin et al. (2023) approached hybrid modeling from a phonological

standpoint by leveraging the wav2vec2 encoder to extract robust acoustic features and map them onto multi-label phonological categories. Their system uses a multi-label CTC loss to classify segments in terms of voicing, place, and manner of articulation, offering fine-grained diagnostic capabilities. The integration of self-supervised pretraining and structured phonological output bridges the gap between deep acoustic modeling and rule-based linguistic interpretation.

Another notable development is the APL (Acoustic, Phonetic, Linguistic) embedding framework proposed by Ye et al. (2022). This model explicitly incorporates three distinct feature types: MFCC-based acoustic vectors, articulatory-phonetic descriptors, and linguistic embeddings derived from phoneme sequences. By aligning and combining these modalities, the system achieves state-of-the-art performance in both detection and diagnosis, with output representations that reflect both physical articulation and phonemic identity. The APL approach exemplifies hybrid philosophy by formalizing the interaction between diverse feature domains.

Extending this line of work into pedagogical applications, Elsayed and Hahn-Powell (2025) developed a CAPT system that integrates acoustic, phonetic, and linguistic representations to provide explicit corrective feedback on learners' grammatical and phonological performance. Built on a fine-tuned ASR backbone, their framework fuses waveform-derived embeddings with symbolic linguistic cues to identify both segmental and suprasegmental deviations. Evaluations in classroom contexts demonstrated significant gains in pronunciation accuracy and learner engagement, underscoring the pedagogical value of hybrid architectures that connect low-level acoustic modeling with linguistically interpretable feedback.

Finally, Zhu et al. (2023) advanced this direction by proposing a feature fusion architecture that uses attention mechanisms to learn the interdependencies among acoustic, phonetic, and linguistic cues. Trained on corpora such as TIMIT and L2-ARCTIC, their model dynamically weights different modalities based on context, allowing it to adaptively prioritize relevant features for each utterance. This flexibility not only improves accuracy but also enhances generalization to diverse speaker profiles and error types.

The growing prevalence of hybrid approaches in pronunciation error detection reflects a strategic convergence of strengths drawn from acoustic, phonetic, and linguistic representations. These systems demonstrate that combining signal-level precision with symbolic interpretability can significantly enhance both detection accuracy and pedagogical relevance. By fusing multiple feature types, whether through end-to-end architectures, attention-based fusion, or augmentation strategies, hybrid models can capture nuanced pronunciation deviations while preserving the structured feedback necessary for learner improvement. However, this complexity often introduces challenges in training and deployment, such as the need for large, annotated datasets, increased computational demands, and difficulties in interpreting the internal mechanics of deep models. Some architectures may obscure the traceability of feedback or rely heavily on phoneme alignment and transcript quality, making them

sensitive to upstream recognition errors. Despite these constraints, hybrid models stand out for their adaptability and comprehensive scope, offering a compelling framework for future PED systems that aim to balance technical performance with educational effectiveness.

Table 4. Summary of hybrid PED approaches

Study	Features	Modeling technique	Strengths	Limitations
Zhang et al. (2020)	Acoustic features + forced alignment + confidence scores	Hybrid CTC/ attention with offline classifier	Multiple scoring methods, phoneme-level feedback.	Multi-stage pipeline, higher computational load.
Lo et al. (2020)	Raw waveforms + latent phonetic info	End-to-end CNN-RNN with attention	Avoids explicit alignment, learns error patterns directly.	Feedback can be opaque, data-hungry
Lo et al. (2021)	Acoustic features + knowledge distillation from DNN-HMM	End-to-end model with label/input augmentation	Combines phonetic info with E2E approach, robust in low-resource setups.	Complex training, tough to interpret intermediate states.
Chilaka (2022)	CTC + acoustic features + phoneme/ character attention	CTC-attention hybrid with linguistic supervision	Captures structured errors, Interpretable.	Requires alignment or transcript supervision, moderate complexity.
Ye et al. (2022)	MFCC, articulatory-phonetic, linguistic embeddings	APL fusion	Multi-level cues, detailed detection & diagnosis.	Requires multiple feature streams and higher system complexity
Shahin et al. (2023)	Self-supervised wav2vec2 + phonological categories	Multi-label CTC classification	Fine-grained phonological feedback is robust under noisy conditions.	Extended training time and CTC constraints limit flexible feedback.
Anwar et al. (2022)	Acoustic features + phoneme supervision	Non-autoregressive dual model	Fast inference, symbolic and acoustic comparison, scalable.	Needs quality phoneme annotations; assumes fixed tokenization.
Elsayed & Hahn-Powell (2025)	Acoustic, phonetic, and linguistic representations	ASR-based CAPT system with multimodal feature fusion	Provided explicit corrective feedback.	Requires extensive model training.
Zhu et al. (2023)	Acoustic, phonetic, and linguistic cues + attention fusion	Attention-based fusion network	Context-aware fusion, high generalization, adaptable.	Training intensive, black box feature interaction

Comparative analysis of feature-based PED aapproaches

Table 5. Comparative summary of feature-based PED approaches

Aspect	Features used	Modeling techniques	Evaluation	Practical use
Acoustic-based	Frame-level MFCCs, PLPs, pitch, energy, spectrograms, or raw waveforms.	SVMs, DNNs, anomaly detection, CNN- RNN architectures using raw/frame-level features.	Evaluated via acoustic similarity measures.	Easy to scale and run in real-time, suitable for mobile or low-resource settings. Good for initial screening and beginner practice, but it provides limited detailed feedback.
Acoustic Phonetic	Acoustic features aligned with phoneme-level units.	GMM-HMM, LDA, DBN-HMM, posterior-based scoring systems.	GOP score, PER, and phoneme-level accuracy for evaluation.	Requires annotated corpora; sensitive to noise but provides phoneme-level error info. Helpful in mid-level CAPT, offering targeted articulatory support; needs careful setup for forced alignment.
Linguistic-based	Phoneme/ word sequences, stress patterns, symbolic rules, extended pronunciation lexicons.	ASR comparison, edit distance, rule-based approaches, template matching.	Typically uses edit distance, match rate, and symbolic scoring.	Highly interpretable errors but depends on robust speech recognition. Generally used in offline/ classroom contexts for structured curricula and phonological error analysis.
Hybrid	Combines acoustic, phonetic, and linguistic cues.	CTC-attention hybrids, transformer encoders, multi-task learning, wav2vec2, feature fusion networks.	mispronunciation F1, PER, alignment accuracy.	High resource demands but yields in-depth, multi-level feedback. Ideal for adaptive tutoring, personalized feedback, and real-time AI-driven platforms.

Discussion and future directions

This survey demonstrates that how a system represents, and processes learner speech fundamentally shapes its overall performance and educational impact in PED. Many methods emphasize raw signal-level data, leveraging standard classifiers or DNNs to analyze features such as frequency distributions, energy, or spectral patterns. These models often show promising speed and scalability, making them a popular choice for real-time or mobile applications with limited resources. Studies using purely signal-driven architectures have indeed achieved efficient mispronunciation screening (Byun & van der Haar, 2019; Chhabra et al., 2023), but the feedback they provide is typically numeric or binary. As a result, learners may not understand which phoneme or prosodic cue caused an error, limiting the system's instructional value.

In contrast, some approaches link the acoustic stream to more explicit

phoneme-level cues or articulatory labels, allowing finer-grained error localization. By incorporating alignment scores, posterior probabilities, or articulatory indicators like formant paths, these systems can yield feedback that correlates closely with expert judgments (Witt & Young, 2000; Jenne & Vu, 2019; Qian et al., 2012). Recent methods employ spectrogram-based CNNs, augmentation techniques such as mixup, or direct waveform modeling to improve resilience to noise and speaker variability (Yan & Chen, 2021; Do et al., 2024). However, these alignment-heavy solutions can be sensitive to transcription errors and often require well-annotated corpora for training. While they offer substantial pedagogical benefits, specifically by pointing out which phoneme is mispronounced, they may be less adaptable to under-resourced languages or spontaneous, accented speech.

Still linguistic methods rely on rule-based symbolic representations, focusing on the linguistic structure of L2 speech. By analyzing phoneme sequences, syllable boundaries, or stress patterns, such systems employ edit-distance or confusion-based scoring to detect systematic deviations rooted in learners' first languages. This symbolic feedback can be highly interpretable: learners see clear rules for substitutions or deletions (Menzel et al., 2000; Helen Meng et al., 2007). Yet, accurate alignment and recognition remain prerequisites, and subtle articulatory deviations or prosodic cues may be overlooked if they do not manifest as discrete phoneme mismatches. Moreover, any noise or accent shifts in the learner's production can reduce the reliability of symbolic detection.

Increasingly, researchers combine multiple streams – acoustic, phonetic, and linguistic- within unified frameworks. These "hybrid" models harness the efficiency of raw acoustic analyses while also offering phoneme-level or symbolic detail. Advanced architectures, such as CTC-attention hybrids, transformer encoders, and self-supervised representations such as wav2vec2, can simultaneously interpret raw waveforms, phoneme posteriors, and symbolic sequences (Chilaka, 2022; Ye et al., 2022; Shahin et al., 2023). This comprehensive approach supports multi-level feedback that pinpoints errors in terms of specific articulatory movements or phoneme substitutions. However, these models can require large amounts of training data and significant computational power, making them less feasible for purely offline or lower-resource environments.

Despite notable progress across these different design choices, unresolved questions persist. Many systems still rely on phone error rate or likelihood-based scores that may not capture overall communicative success or educational improvement. Researchers are beginning to propose new metrics that incorporate pedagogical needs while also addressing cross-linguistic variability and learner diversity. In parallel, self-supervised learning has gained traction, as pretraining on massive unlabeled corpora can reduce dependence on scarce annotated data (Shahin et al., 2023). Multimodal approaches that include visual or textual cues can further illuminate articulatory targets, enhancing learner comprehension and motivation.

From a CALL perspective, the effectiveness of PED systems depends not only on detection accuracy but also on how feedback supports learning. Studies

show that explicit, interpretable feedback enhances learner engagement and improvement, suggesting that feature-based PED models should integrate pedagogical design with technological precision (Amrate & Tsai, 2024; Elsayed & Hahn-Powell, 2025).

Ultimately, no single approach excels at all dimensions of speed, interpretability, language adaptability, and real-time feedback without trade-offs. Systems relying primarily on raw signal-level data excel at efficiency but offer limited instructional guidance, whereas symbolic or hybrid models provide deeper insights but require extensive annotation, computational resources, or robust alignment. Future work will likely focus on bridging these gaps by developing data-efficient architectures, adaptive feedback loops, and evaluation frameworks that reflect authentic improvements in learner pronunciation. By addressing these challenges collaboratively and drawing on insights from speech technology, applied linguistics, and pedagogy, next-generation PED systems could become more accessible, intelligible, and effective for learners worldwide.

Conclusion

This survey presented a feature-centric taxonomy of PED systems, emphasizing how input representations acoustic, acoustic-phonetic, linguistic, and hybrid profoundly influence diagnostic precision, interpretability, and real-world applicability. Each category offers distinct benefits, from the scalability of purely acoustic approaches to the granular feedback of acoustic-phonetic methods, the symbolic clarity of linguistic-based models, and the comprehensive insights afforded by hybrid systems. Yet, each also poses trade-offs in terms of resource requirements, generalizability, and the depth of corrective feedback.

A key lesson is that effective PED extends beyond mere detection accuracy. Systems must provide actionable, learner-centered feedback and adapt to varying linguistic backgrounds and proficiency levels. Promising advances in self-supervised learning and multi-modal integration show potential for improving robustness and personalization. With sustained collaboration between educators and technologists, future PED solutions can better balance efficiency, interpretability, and pedagogical impact, ultimately enhancing second language pronunciation acquisition.

Use of generative AI

This manuscript made use of ChatGPT, Gemini to assist in locating relevant literature, gathering reference information, paraphrasing sections of text for clarity, refining academic language, and ensuring consistency in citation formatting.

No AI tools were used for the generation of research hypotheses or the writing of conclusions. The final manuscript was authored and reviewed by the listed authors without additional AI input.

References

Amrate, M., & Tsai, P.-H. (2024). Computer-assisted pronunciation training: A systematic review. *ReCALL, 36*(1), 22–42.

Anwar, M., Alatiyyah, M. H., & Mridha, M. F. (2022). Non-autoregressive end-to-end neural modeling for automatic pronunciation error detection. *Applied Sciences, 13,* 109–109.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics, 32,* 233–250. https://doi.org/10.1016/S0095-4470(03)00036-6

Bartelds, M., Richter, C., Liberman, M., & Wieling, M. (2020). A new acoustic-based pronunciation distance measure. *Frontiers in Artificial Intelligence, 3,* 39.

Byun, J., & van der Haar, D. (2019). Pronunciation detection for foreign language learning using MFCC and SVM. In Information Science and Applications 2018: ICISA 2018 (pp. 327–335).

Chhabra, P., Chhillar, S., Tanwar, R., Verma, M., & Indra, G. (2023). Mispronunciation detection using feature learning. In *International Conference on Computing and Communication Networks* (pp. 307–316).

Chilaka, S. (2022). Improved mispronunciation detection system using a hybrid CTC-ATT based approach for L2 English speakers. https://doi.org/10.48550/ARXIV.2201.10198

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39,* 379–397. https://doi.org/10.2307/3588486

Do, H. Lee, W., & Lee, G. G. (2024). Acoustic feature mixup for balanced multi-aspect pronunciation assessment. In *Proceedings of the 25th Interspeech Conference 2024* (pp. 312–316). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2024-2498

Ehrlich, S., & Avery, P. (2013, May). *Teaching American English pronunciation: Oxford handbooks for language teachers.* Oxford University Press.

Elsayed, I., & Hahn-Powell, G. (2025). Computer-assisted pronunciation training for foreign language learning of grammatical features. *Language Learning Technology, 29*(1), 1–20.

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

Flege, J. E. (1995, January). Second language speech learning: Theory, findings and problems. Retrieved from

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "l" and "r". *Neuropsychologia, 9,* 317–323.

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL, 15*(1), 3–20. 10.1017/S0958344003000211.

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

Jenne, S., & Vu, N. T. (2019). Multimodal articulation-based pronunciation error detection with spectrogram and acoustic features. *Interspeech* 2019, 3549–3553. https://doi.org/10.21437/interspeech.2019-1677

Lee, A., & Glass, J. (2015). Mispronunciation detection without nonnative training data. *Interspeech* 2015, 643–647.

Lee, C. H., & Hsieh, M.-H. (2017). Effects of using an automatic speech recognition based pronunciation learning system on L2 learners' pronunciation development. *Computer Assisted Language Learning, 30*(4), 401–424. https://doi.org/10.1080/09588221.2017.1327758

Lo, T., Sung, Y., & Chen, B. (2021). *Improving end-to-end modeling for mispronunciation detection with effective augmentation mechanisms.* 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1049–1055.

Lo, T.-H., Weng, S.-Y., Chang, H.-J., Chen, B. (2020). An effective end-to-end modeling approach for mispronunciation detection. In *Proceedings of Interspeech* 2020 (pp. 3027–3031). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2020-1605

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 and Understanding* (ASRU). https://doi.org/10.1109/ASRU.2007.4430152

Menzel, W., Herron, D., Bonaventura, P., & Morton, R. (2000). Automatic detection and correction of non-native English pronunciations.

Metruk, R. (2024). Mobile-assisted language learning and pronunciation instruction: A systematic literature review. *Education and Information Technologies, 29*, 16255–16282.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38,* 289–306.

Neri, A., Cucchiarini, C., & Strik, H. (2006). ASR-based corrective feedback on pronunciation: Does it really work? *Interspeech* 2006, paper 1372. https://doi.org/10.21437/interspeech.2006-543

Qian, X., Meng, H., & Soong, F. K. (2012). The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. *Interspeech* 2012, 775–778.

Shahin, M., & Ahmed, B. (2019). Anomaly detection based pronunciation verification approach using speech attribute features. *Speech Communication, 111*, 29–43.

Shahin, M., Epps, J., & Ahmed, B. (2023). *Phonological level wav2vec2-based mispronunciation detection and diagnosis method.* arXiv.org

Stockwell, G. (2012). *Computer-assisted language learning: Diversity in research and practice.* Cambridge University Press.

Strik, H., Cucchiarini, C., & Truong, K. P. (2009). Automatic pronunciation error detection and feedback generation for CALL applications. In H. Höge & M. Jansche (Eds.), *Speech and language processing for assistive technologies* (pp. 188–198). Springer. https://doi.org/10.1007/978-3-319-20609-7_17

Strik, H., Truong, K., de Wet, F., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication, 51*, 845–852.

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., Neri, A., Cucchiarini, C., & Strik, H. (2004). Automatic pronunciation error detection: An acoustic-phonetic approach (pp. 135–138).

Wennerstrom, A. (2001, November). *The music of everyday speech.* Oxford University Press.

Witt, S., & Young, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication, 30,* 95–108. https://doi.org/10.1016/S0167-6393(99)00044-8

Yağız, O., Kaya, F., & Ötügen, R. (2024). The effectiveness of L2 pronunciation instruction: A critical systematic review of the intervention studies. *The Literacy Trek, 10*(1), 21–41. https://doi.org/10.47216/literacytrek.1487243

Yan, B., & Chen, B. (2021). End-to-end mispronunciation detection and diagnosis from raw waveforms. In *2021 29th European Signal Processing Conference (EUSIPCO).*

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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)* (pp. 6827–6831). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746135

Yonesaka, S. M. (2017). Learner perceptions of online peer pronunciation feedback through P-Check. *The JALT CALL Journal, 13*(1), 25–91. https://doi.org/10.29140/jaltcall.v13n1.j210

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

Zhu, C., Wumaier, A., Wei, D., Fan, Z., Yang, J., Yu, H., Kadeer, Z., & Wang, L. (2023). Pronunciation error detection model based on feature fusion. Speech Communication, 156, 103009.