Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17899-x

Mispronunciation detection and diagnosis using deep neural networks: a systematic review

Meriem Lounis[1,2] · **Bilal Dendani**[1,2] · **Halima Bahi**[1,2]

Received: 17 September 2022 / Revised: 17 September 2023 / Accepted: 13 December 2023

Abstract
The increased need for foreign language learning, along with advances in speech technology have heightened interest in computer-assisted pronunciation teaching (CAPT) applications. Herein, the automatic diagnosis of pronunciation errors is essential, it allows language learners to identify their mispronunciations and thus improve their oral skills. Meanwhile, the emergence of deep learning algorithms for speech processing led to the use of deep neural networks at several stages of the mispronunciation detection and diagnosis process. Therefore, an overview of the state-of-the-art in deep learning algorithms for mispronunciation diagnosis is needed, for which we performed a systematic literature review. This study aims to provide an overview of the recent use of deep neural networks for mispronunciation detection and diagnosis (MDD). A thorough statistical analysis is provided in this review which was conducted by extracting specific information from 53 papers published between the years 2015 and 2023. This review indicates that the diagnosis of pronunciation errors is a highly active area of research. Quite a few deep learning models and approaches have been proposed in this area, but there are still some important open issues and limitations to be addressed in future works.

Keywords Computer-Assisted Pronunciation Training (CAPT) · Mispronunciation Detection and Diagnosis (MDD) · Deep Learning (DL) · Deep Neural Networks (DNN) · Systematic review

✉ Bilal Dendani
bilal.dendani@univ-annaba.dz

Meriem Lounis
meriem.lounis@univ-annaba.org

Halima Bahi
halima.bahi@univ-annaba.dz

[1] Computer Science Department, University Badji Mokhtar, Annaba, Algeria

[2] LISCO Laboratory, University Badji Mokhtar, Annaba, Algeria

Published online: 09 January 2024

1 Introduction

Globalization and the virtualization of frontiers have led to an increased demand for learning foreign languages. Speech being the most intuitive and popular mode of communication, pronunciation learning takes a considerable place. Particularly, computer-assisted pronunciation teaching (CAPT) attracts much attention; CAPT is a part of computer-assisted language learning (CALL) systems that deals with pronunciation. Herein, the automatic pronunciation assessment aims to automatically assess pronunciation as an expert would do; furthermore, the automated assessment process "reduce[s] … the need of human intervention significantly saving cost and time" [1, p.1]. Pronunciation assessment aims to provide the learners with informative feedback that pinpoints the pronunciation errors and allows them to improve their outcomes; the feedback is possible through mispronunciation detection and diagnosis (MDD).

From the beginning, pronunciation assessment was tightly related to speech recognition technology [2]. For a long time, automatic speech recognition (ASR) technology based on hidden Markov models (HMMs) was the key to automatic pronunciation assessment and error detection as well [3–6]. Herein, the pronunciation to assess is force-aligned to a reference HMM, and "the system outputs a score representing how close is the incoming speech to the correct pronunciation" [7, p.2]. A badly pronounced utterance is expected to generate low scores, while a native-like utterance would better fit the model and, hence, generate higher scores. The force-alignment stage provides a collection of confidence measures (CM), such as log-likelihood and phone duration [8], which serve to compute other scores such as the well-known goodness of pronunciation (GOP) [6].

In the earlier systems, the diagnosis was handled as a classification task where the classes stand for the set of possible sounds that may be pronounced given a target one [9], the aforementioned GOP is the most used score to separate the classes. Later, the extended recognition network (ERN) approach arises as an alternative to the classification approach. An extended recognition network is a finite state transducer (FST) that includes both the canonical pronunciation and the common mispronunciations [10]. The performances of both classification and ERN approaches depend on the ability to provide high coverage of the possible mispronunciation patterns, which is often difficult to guarantee due to unforeseen errors, and multiple background influences, especially for multilingual learners. Therefore, the unsupervised pronunciation error detection approach appeared, herein, possible unknown errors are discovered during the clustering [11]; the approach is suitable for low-resource languages and in the context of personalized learning as well [12].

On the other side, the MDD tasks can be implemented either in one-pass or two-pass. The first and most popular approach is the one-pass pronunciation error detection and diagnosis with an ERN. "On the other hand, the two-pass framework detects the places where there are possible errors in the first pass. In the second pass, phone loop recognition is conducted at the problematic places to identify the actual error types." [13, p. 3].

Despite numerous pieces of research linked to approaches that implement the MDD tasks in CAPT systems, MDD is still hard to implement. Mainly, due to the scarcity of large-scale linguistic resources such as the lack of nonnative speech data, and the lack of human annotations. Lately, the emergence of deep learning (DL) algorithms known to discover and model complex relationships among data boosted the research in pronunciation assessment and pronunciation error diagnosis. In this context, deep neural networks (DNNs) which are neural networks with many hidden layers have been used as feature extractors, classifiers, or end-to-end (E2E) architectures.

This review reports the recent works that deal with mispronunciation detection and diagnosis using DNNs. For this purpose, a secondary study is developed following [14] guidelines to perform a systematic literature review. This review aims to identify, evaluate, and interpret all available research on mispronunciation diagnosis based on DNNs. For the first time, the identified number of papers was 403 papers, published between 2015 and 2023 (August), nevertheless, after applying the inclusion/exclusion criteria, and the quality assessment rules, only 53 primary study papers were included in our summary. The research questions were answered by extracting suitable information from these papers, in the form of statistical representations using tables and figures. The presented results are intended to show the trend of research in this area, underline challenges, and focus on new opportunities.

The article is organized as follows. Section 2 summarizes the related work. Section 3 presents the methodology used to conduct this review, including the research questions, and the inclusion/exclusion criteria. In Section 4, the review results are presented including the answers to the research questions. In Sect. 5, the results are discussed and the responses to the research questions are summarized. Section 6 concludes the paper.

2 Related work

The mispronunciation diagnosis is a milestone in the language learning process; it allows the production of informative feedback that helps learners correct their mistakes. However, no review has been exclusively interested in approaches and algorithms used for this purpose. Meanwhile, few reviews reported the works that used automatic speech recognition for pronunciation assessment purposes. First, Neri et al. [15] described the ASR-based CAPT systems as a sequence of five phases. The first is speech recognition, the authors considered this phase as the “most important phase, as the subsequent phases depend on the accuracy of this one” [15, p. 1]. The second phase is the scoring, the speech recognition stage outputs a collection of confidence measures [5] that represent how close is the incoming speech to the canonical pronunciation, such as the log-likelihood score. The error detection is the third phase which is mainly based on the confidence measures. The fourth phase is the error diagnosis. Herein, the system identifies the type of pronunciation errors made by the learner. The last phase is the feedback which allows the learner to benefit from the information provided by the previous phases. In this review, the authors have reported several problems related to the first phase that were raised in the subsequent ones. In 2009, Eskenazi [3] reviewed the works in spoken language technology for education, from its beginning in the late 1980s. In particular, the author stated that automatic speech recognition was the main technology used for language learning systems. In 2012, Witt [16] published a state-of-the-art on automatic error detection from its beginning in 1990 until 2012.

In 2016, Chen and Li [4] reviewed the approaches used in CAPT systems including mispronunciation detection and diagnosis from 2012 to 2016. The authors proposed a classification depending on pronunciation patterns either phonetic errors or prosodic errors. From the phonetic errors point of view, the authors reviewed the scoring-based approach, the classification approach, the ERN approach, and the unsupervised error discovery approach. Among the strategies for improving phonetic error detection, the authors emphasized the contribution of DNNs to improve the confidence measures (including the GOP) when they are used instead of Gaussian mixture models (GMMs) in the acoustic modeling of phonemes. The HMMs model the temporal progression of the speech signal, from one state to another, while the GMMs model the observations within the state. Indeed, “deep neural networks … have been shown to outperform GMMs on a variety of speech recognition benchmarks, sometimes by a large margin.” [17, p. 82].

Lately, Agarwal and Chakraborty [18] reviewed the available CAPT tools for English learners. This review is structured according to the classification of CAPT systems adopted by the authors, namely: visual simulation-based systems, game-based systems, comparative phonetics-based systems, and artificial neural network-based systems. Even if English is an extensive resource language, and “several CAPT systems have been developed and used successfully for teaching English pronunciation” [18, p. 3741], the authors concluded that “more effort is needed on the development of CAPT systems”. [18, p. 3741].

As seen above, there have been only a few comprehensive reviews that studied pronunciation assessment, and none exclusively the mispronunciation diagnosis. Moreover, none of them reviewed and discussed the use of deep learning algorithms in mispronunciation diagnosis. This review aims to address the gap previously raised. Thus, it focuses exclusively on the diagnosis of pronunciation errors. The automatic pronunciation error diagnosis is the most difficult task to implement since it requires a lot of linguistic resources and is the most important for the production of informative/corrective feedback. We are also interested in the contribution of deep learning algorithms in error pronunciation diagnosis, firstly because this aspect has been little or not at all discussed in the previous second studies, and because, in recent years, these algorithms have shown great performances in feature extraction and classification as well as in knowledge discovery.

3 Methodology

A systematic literature review (SLR) is a methodological review of research results; it involves several discrete activities. This study was conducted according to the guidelines described in [14]; the guide summarizes the stages in a systematic review into three main phases: planning the review, conducting the review, and reporting the review. In the planning stage, based on the review goals, research questions are determined as well as the search protocol. During the conducting stage, the research papers are selected, and data extracted from the research papers are synthesized. The reporting stage consists of writing and disseminating the results of the review to stakeholders. The review protocol that was followed is detailed below.

3.1 Research questions

This review aims to identify and examine the articles that implement DNNs in the area of mispronunciation diagnosis, and it also aims to identify trends, challenges, and possible opportunities from the existing research. Based on that, the following research questions (RQ) were identified:

RQ1: What are the different types of papers that were included in this study?
RQ2: What are the different corpora/languages identified in the selected papers?
RQ3: What are the pronunciation error patterns identified in the selected papers?
RQ4: What are the different MDD approaches identified in the selected papers?
RQ5: What deep neural network algorithms have been used in the selected papers?
RQ6: What evaluation techniques were used in the selected papers?

3.2 Search strategy

The search terms used in the study were derived based on main terms from research questions and terms obtained from pertinent papers. As the current SLR is twofold, the following terms were identified: “deep learning”, “deep neural network”, “deep neural networks”, and “DNN” for the deep learning side, and “mispronunciation detection and diagnosis”, “pronunciation error detection and diagnosis” for the MDD side.

The search terms were combined into one meta-search term. Alternative synonyms were combined using the disjunction operator “OR” and major terms were linked using the “AND” operator. The search string is as follows: ("deep learning" OR "deep neural network" OR "deep neural networks" OR “DNN”) AND ("mispronunciation detection and diagnosis" OR "pronunciation error detection and diagnosis").

The defined keywords were used in six databases. Namely: IEEE Explorer, Science Direct, Springer, ACM Digital Library, Scopus, and Google Scholar.

3.3 Study selection

Based on the search terms, 403 papers were obtained. Their citations were then imported into a worksheet. After that, selection and filtration were conducted according to the following steps:

Step 1: remove review papers from the list of papers.
Step 2: apply inclusion/exclusion criteria to keep only relevant papers.
Step 3: remove all duplicate research papers that were obtained from different digital libraries.
Step 4: apply quality assessment rules to select papers with the highest quality.

Once, the review papers are removed from the list, inclusion/exclusion criteria are applied. The used inclusion/exclusion criteria in this review paper are defined below:

Inclusion criteria:

Papers that use DNNs for mispronunciation diagnosis for pronunciation learning/teaching purposes,
Written in English,
Published between 2015 and 2023.

Exclusion criteria:

Review articles,
Not peer-reviewed papers.

Finally, one quality assessment rule (QAR) was applied to select the final list of papers; it consists of considering only papers that have a clear methodology, results, and discussion sections. The source selection process is shown in Fig. 1.

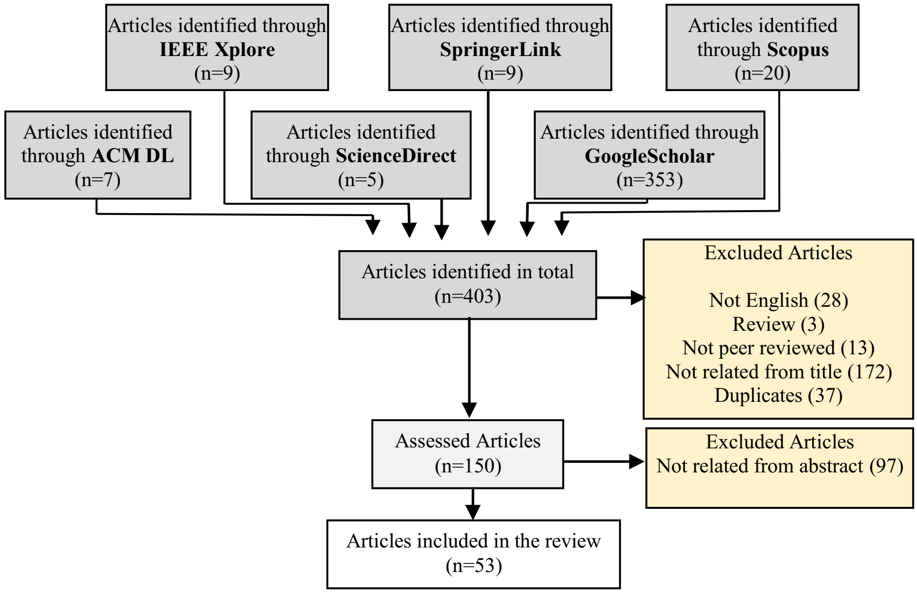
Articles identified through **IEEE Xplore** (n=9)
Articles identified through **SpringerLink** (n=9)
Articles identified through **Scopus** (n=20)
Articles identified through **ACM DL** (n=7)
Articles identified through **ScienceDirect** (n=5)
Articles identified through **GoogleScholar** (n=353)
Articles identified in total (n=403)
Excluded Articles
Not English (28)
Review (3)
Not peer reviewed (13)
Not related from title (172)
Duplicates (37)
Assessed Articles (n=150)
Excluded Articles
Not related from abstract (97)
Articles included in the review (n=53)

Fig. 1 Source selection process from literature dataases

3.4 Data extraction

Finally, 53 papers were extracted, and these papers were studied in depth. Table 1 presents the selected studies that met the selection criteria. The relevant data were extracted by reading all the selected studies in full, and the extracted information was used to answer the research questions.

4 Data synthesis and result analysis

In this section, we summarize the studies obtained as a result of the review process. The results from the research questions were reported and presented as quantitative data and used to develop a statistical comparison between the different findings for each research question. We analyze these results to identify the trends, underline challenges, show opportunities, and thus provide recommendations for future work.

4.1 Year-wise distribution of the selected papers (RQ1)

The selected 53 papers fall into four main categories, which are: conference papers, journal papers, workshop papers, and book chapters. Figure 2 shows the year-wise distribution of the extracted papers and provides their distribution among these categories. The majority of the papers with 58% are from conferences, and 34% were extracted from journals.

Table 1 Database-wise distribution of the selected papers

Base	Year	Ref	Title
ACM DL	2020	[13]	Cross-Lingual Transfer Learning of Non-Native Acoustic Modeling for Pronunciation Error Detection and Diagnosis
	2019	[19]	The Use of SDAE in Noisy English Mispronunciation Detection and Diagnosis towards Application in Mobile Learning
Elsevier	2018	[20]	Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks
	2017	[21]	Intonation classification for L2 English speech using multi-distribution deep neural networks
Google Scholar	2023	[22]	Arabic mispronunciation recognition system using LSTM Network
		[23]	Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis
		[24]	End-to-End Mispronunciation Detection and Diagnosis Using Transfer Learning
		[25]	Multi-Feature and Multi-Modal Mispronunciation Detection and Diagnosis Method Based on the Squeezeformer Encoder
		[26]	Peppanet: effective mispronunciation detection and diagnosis leveraging phonetic, phonological, and acoustic cues
		[27]	Phonetic RNN-Transducer for Mispronunciation Diagnosis
	2022	[28]	An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings
		[29]	Masked acoustic unit for mispronunciation detection and correction
		[30]	Maximum f1-score training for end-to-end mispronunciation detection and diagnosis of l2 english speech
		[31]	Mispronunciation Detection and Diagnosis with Articulatory-Level Feedback Generation for Non-Native Arabic Speech
		[32]	Self-Supervised Pre-Trained Speech Representation Based End-to-End Mispronunciation Detection and Diagnosis of Mandarin
	2021	[33]	A computer-aided speech analytics approach for pronunciation feedback using deep feature clustering
		[34]	Automatic Detection of Word-Level Reading Errors in Non-native English Speech Based on ASR Output
		[35]	Detection of Mispronunciation in Non-native Speech Using Acoustic Model and Convolutional Recurrent Neural Networks

Table 1 (continued)

Base	Year	Ref	Title
		[36]	End-to-end mispronunciation detection and diagnosis from raw waveforms
		[37]	Improving Mispronunciation Detection of Mandarin for Tibetan Students Based on the End-To-End Speech Recognition Model
		[38]	Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning
		[39]	Text-conditioned transformer for automatic pronunciation error detection
		[40]	Transformer Based End-to-End Mispronunciation Detection and Diagnosis
	2020	[41]	L2 Mispronunciation Verification Based on Acoustic Phone Embedding and Siamese Networks
		[42]	SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis
	2019	[43]	A Study on Mispronunciation Detection Based on Fine-grained Speech Attribute
		[44]	CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis
		[45]	Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and blstm-based deep tone models
		[46]	Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for Arabic phonemes
		[47]	Pronunciation Erroneous Tendency Detection with Combination of Convolutional Neural Network and Long Short-Term Memory
	2018	[48]	Applying multitask learning to acoustic-phonemic model for mispronunciation detection and diagnosis in l2 english speech
		[49]	Unsupervised discovery of an extended phoneme set in l2 english speech for mispronunciation detection and diagnosis
	2017	[50]	A study of automatic annotation of pets with articulatory features
		[51]	Articulatory Modeling for Pronunciation Error Detection without Non-Native Training Data Based on DNN Transfer Learning
		[52]	Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data
		[53]	Improving pronunciation erroneous tendency detection with convolutional long short-term memory

Table 1 (continued)

Base	Year	Ref	Title
		[54]	Mispronunciation Diagnosis of L2 English at Articulatory Level Using Articulatory Goodness-Of-Pronunciation Features
		[55]	Transfer Learning based Non-native Acoustic Modeling for Pronunciation Error Detection
	2016	[56]	Context Aware Mispronunciation Detection for Mandarin Pronunciation Training
		[57]	Multi-lingual and multi-task DNN learning for articulatory error detection
	2015	[58]	A study on robust detection of pronunciation erroneous tendency based on deep neural network
IEEE	2022	[59]	Exploring Non-Autoregressive End-to-End Neural Modeling for English Mispronunciation Detection and Diagnosis
	2021	[60]	Mispronunciation Detection and Diagnosis for Mandarin Accented English Speech
	2018	[61]	Integrating Articulatory Features into Acoustic-Phonemic Model for Mispronunciation Detection and Diagnosis in L2 English Speech
	2017	[62]	Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks
	2016	[63]	Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling
Scopus	2021	[64]	A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis
	2018	[65]	Unsupervised Discovery of Non-native Phonetic Patterns in L2 English Speech for Mispronunciation Detection and Diagnosis
	2017	[66]	Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models
		[67]	Phonological Feature Based Mispronunciation Detection and Diagnosis Using Multi-Task dnns and Active Learning
	2016	[68]	Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-Guided and Data-Driven Decision Trees
	2015	[69]	An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech
Springer	2019	[70]	Mandarin Chinese Mispronunciation Detection and Diagnosis Leveraging Deep Neural Network Based Acoustic Modeling and Training Techniques

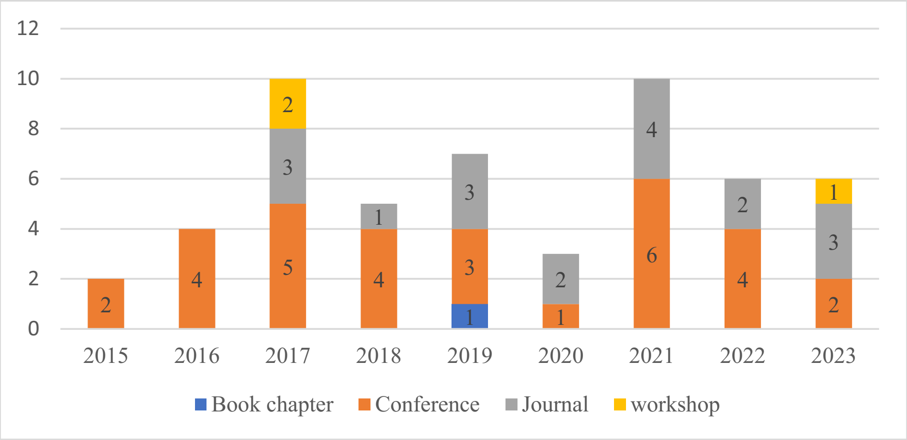

Fig. 2 Year-wise distribution of the selected papers

Figure 2 shows that early work on the use of DNNs in MDD was almost exclusively presented at conferences. Over time, the maturity of the subject has meant that more publications are made by journals. Table 2 presents the distribution of published papers among conferences, journals, workshops, and book chapters.

4.2 Speech corpora (RQ2)

The condition for the development of deep learning models is the availability of large datasets. While native speech corpora are used to develop ASR-based systems for a given language, the development of MDD systems requires the availability of nonnative corpora dedicated to CAPT purposes. Indeed, the pronunciation acquisition by L2 learners is tightly related to their L1 phonological systems. Thus, if the L1 learner's background is known, the pronunciation errors are more likely to be easily detected and diagnosed.

In the selected papers, many native speech corpora such as TIMIT were used to train the ASR models. Meanwhile, it is worthy to notice the scarcity of CAPT-dedicated nonnative speech corpora. More details are shown in Table 3 about the availability of those corpora; Table 3 also shows that English and Mandarin are among the high-resource languages in the MDD topic.

Table 3 shows that most corpora are public, which contributes to the production of more research and more tools for English and Mandarin. Meanwhile, Arabic corpora are private which slows the evolution of work for Arabic and similarly for other low-resource languages. Indeed, deep learning models require a huge amount of data for the training.

On the other hand, the trend in language learning is toward ubiquitous systems, where the learner can practice anywhere and at any time, herein, the use of dedicated corpora for the DNN models is crucial. However, most of the available corpora were recorded in quiet rooms such as for L2-Arctic and iCALL. In fact, among all of the selected papers, only one study deals with a noisy environment [19].

Table 2 Distribution of conferences/journals papers

Journal / Conference / Book Chapter/WorkShop	# Selected papers
Book chapter	1
Chinese Language Learning Sciences	1
Conferences	31
ICASSP	11
Interspeech	8
APSIPA	3
Others	9
Journals	18
IEEE Access Applied sciences	3
IEEE/ACM Transactions on Audio, Speech, and Language Processing	3
Speech Communication	2
Applied Sciences	1
Computer Speech and Language	1
IEICE Trans	1
Information	1
International Journal of Asian Language Processing	1
Journal of Physics	1
Journal of Signal Processing Systems	1
Mathematics	1
Multimedia Systems	1
Neural Networks	1
Workshops	3
ISCA	2
IEEE Spoken Language Technology Workshop (SLT)	1
Total	53

4.3 The studied mispronunciation patterns (RQ3)

Pronunciation errors belong into two types: phonemic errors (Ph.), and prosodic errors (Pr.). Figure 3 details the possible patterns of pronunciation errors.

Even if, it was shown that prosodic errors impact intelligibility more than segmental errors [71], Fig. 4 shows that, in the selected studies, segmental errors were much more addressed than suprasegmental ones.

4.4 Deep neural network models and purposes (RQ4 -RQ5)

As already said, the diagnosis is performed as a classification task, through an ERN as an ASR-based model, or in an unsupervised way where errors are discovered. Figure 5 shows the distribution of the extracted papers based on the used approach.

Table 3 Nonnative corpora used in the selected papers

Corpus	#Count
Arabic	4
Private	4
KSU speech corpus	1
L2 Arabic	3
English	28
Private	4
"RA" corpus	1
L2 English	2
Speech Accent Archive pronunciation corpus + Noise-92 corpus	1
Public	24
CU-CHLOE corpus	7
EMA-MAE Corpus	1
ETRI Corpus (Electronics and Telecommunications Research Institute)	1
Interactive Spoken Language Education (ISLE) Dataset	1
L2-Arctic Corpus	12
Supra-CHLOE corpus	2
Mandarin	20
Private	4
L2 Mandarin	2
PSC-Reading dataset	2
Public	16
BLCU Corpus	10
iCALL Corpus	5
Mandarin annotated spoken (MAS) native and non-native Corpus	1
English and Cantonese	1
Private	1
Isolated Word corpus	1
Total	53

As shown in Fig. 5, most of the selected studies fall in the classification approach as MDD is mostly treated as a classification problem, and the unsupervised method becomes a new trend for MDD approaches. The ERN-based method is parsimoniously used due to its difficulty in modeling all the possible mispronounced phone pairs. The classification approach as well as the ERN-based assumes a set of mispronounced phonemes designed based on the L1 background of the learner while the unsupervised approach allows L1 independency.

In this review, DNN models are classified into five categories according to the task they performed, namely: acoustic modeling, feature extraction, classification, ASR, and E2E, Table 4 shows the paper references for each class.

The acoustic models corresponding to phone segments had been commonly parameterized with HMMs and GMMs (GMM–HMM). More recently, with the successful development of deep neural networks, they have been employed instead of GMM for the state emission probabilities estimation in HMMs, leading to DNN–HMM-based acoustic models for ASR purposes, including MDD tasks [69, 72].

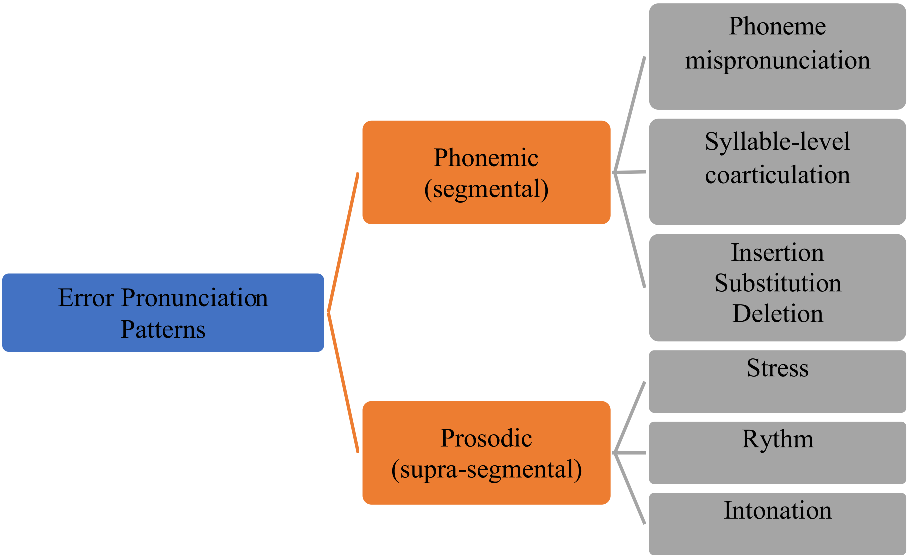
Error Pronunciation Patterns
Phonemic (segmental)
Prosodic (supra-segmental)
Phoneme mispronunciation
Syllable-level coarticulation
Insertion Substitution Deletion
Stress
Rythm
Intonation

Fig. 3 Type of pronunciation errors

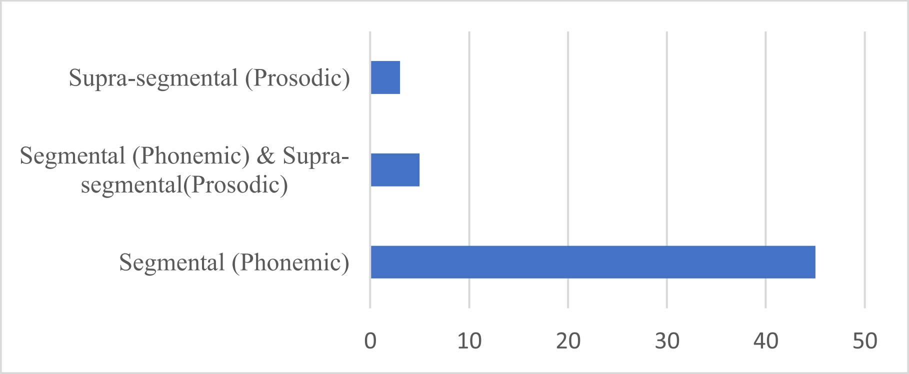

Fig. 4 Distribution of the pronunciation error patterns in the selected studies

MDD is mainly considered as a classification task, therefore, DNN models were used as classifiers and feature extractors as well. Tables 4 and 8 report the studies that used DNN models as classifiers or as feature extractors.

Finally, E2E is a recent trend that aims to apply the DNN-based ASR paradigm to MDD, "which in essence employs a free-phone recognition topology implemented with deep neural networks like connectionist temporal classification (CTC), attention-based model or their hybrids (denoted by hybrid CTT-ATT)"[73, p. 1065].

Tables 5 and 6 report the several DNN models used in the selected studies, either as a standalone model or in a hybrid manner.

As shown in Table 5, the multilayer perceptron is the most used architecture, indeed, it was the first used model in the context of DNN-HMM-based ASR architectures [65]. Later, encoders have been widely used for feature extraction purposes. An encoder is

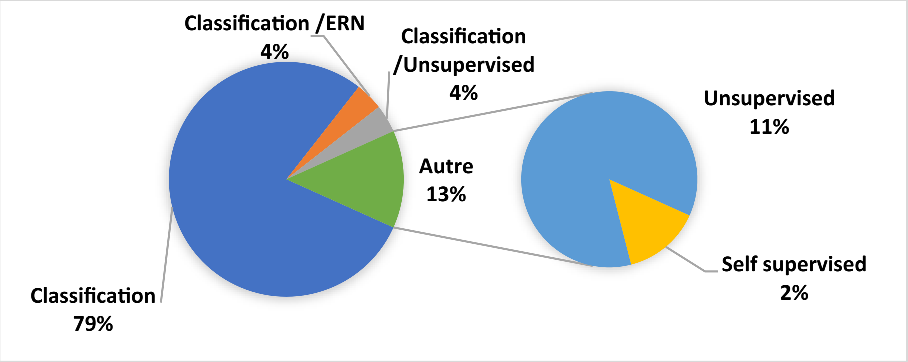

Fig. 5 Distribution of extracted papers over the MDD approaches

a particular architecture where the dimensionality of the input pattern is reduced from one hidden layer to the next, up to a latent layer that captures the essence of the input representation while reducing the space dimensionality [19, 25].

Meanwhile, CNN algorithms and RNN ones have been used in a hybrid way to enhance local feature extraction and to capture temporal dependencies as well [28, 31, 35, 44].

Tables 5 and 6 show the wide use of DNN models; moreover, these tables show the popularity of the recurrent neural network (RNN) algorithm and its derivatives such as the long-short time memory(LSTM), the gated recurrent unit (GRU), etc. RNN performs computations on the time sequence since its current hidden state is dependent on all the previous hidden states. More specifically, it is designed to model time-series signals as well as capture long-term and short-term dependencies between different time steps of the input. Indeed, the CNN-RNN models and their several implementations, in particular, CNN-CTC-ATT are the most used ones.

Finally, it is worth noticing that transformers, which become state-of-the-art in ASR [74] applications, tend to replace CNN-CTC algorithms in favor of better performances [25, 26, 32, 39, 40, 64].

Table 4 Classification of DL algorithms for MDD based on the performed task

Performed task	Key references
Acoustic modelling	[13, 27–29, 31, 34, 35, 38, 43, 45, 48, 49, 51, 52, 54–58, 62, 67, 69, 70]
DL-based ASR	[53, 60]
End-to-End	[23–26, 30, 32, 36, 37, 39–42, 44, 46, 47, 59, 64]
Features extraction	[19, 20, 33, 45, 46, 50, 53, 61, 63, 65, 66, 68]
Classification	[21, 22, 45, 63, 66],

Table 5 Standalone DL models

DL algorithm	# Selected papers
DNN (Multilayer perceptron (MLP), Feedforward neural network (FFNN), autoencoder)	15
Transfer Learning (TL) (with DNN architecture)	7
Multi-task (MT)—DNN	5
Transformer	5
Convolutional neural network (CNN) (AlexNet)	2
Gated recurrent unit (GRU)	2
Long short-term memory (LSTM)	2
Time-delay neural network (TDNN)	2
Bi-directional LSTM (BiLSTM)	1
CNN	1
Graph convolutional network (GCN)	1
Recurrent neural network (RNN)	1
Stack Denoising Autoencoder (SDAE)	1
Siamese CNN	1
Siamese DNN	1
Vector-quantized variational autoencoder (VQ-VAE)	1

Table 6 Hybrid DL models

DL algorithm	# Selected papers
CNN-RNN-CTC (CTC for Connectionist temporal classification)	4
DNN-HMM	4
CNN-LSTM	2
DNN-LSTM	2
CNN-GRU-CTC	1
CNN-TRANSFORMER-CTC	1
encoder-(CNN-RNN)-GRU	1
SincNet (or CNN)-CTC-ATT (ATT for Attention)	1
Squeezeformer-BiLSTM-Transformer	1
Transformer- Wav2Vec2.0	1

4.5 Evaluation techniques (RQ6)

When it comes to evaluating the algorithms for MDD, a few measures must be defined. Fig. 6 shows the evaluation hierarchy issued from [75].

As the MDD is often considered a classification task, almost all selected papers used the related performance metrics. Mainly, we mention: precision, recall, and the F1-score, defined as Eq. 1, Eq. 2, and Eq. 3 respectively.

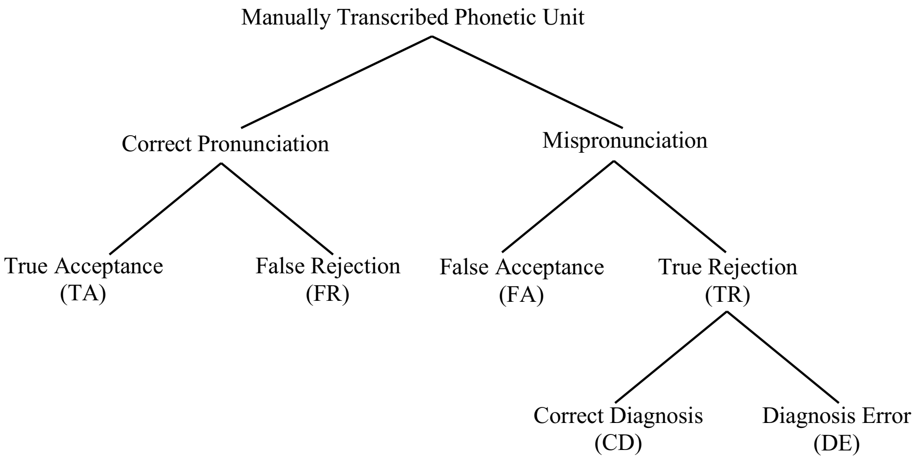

Fig. 6 The hierarchical structure of the mispronunciation detection and diagnosis errors

$$Precision = \frac{TR}{TR + FR} \tag{1}$$

$$Recall = \frac{TR}{TR + FA} \tag{2}$$

$$F1 - score = \frac{2(Precision * Recall)}{Precision + Recall} \tag{3}$$

Meanwhile, the detection performances are measured using the false rejection rate (FRR), the false acceptance rate (FAR), and the detection accuracy (DetAcc), defined as Eq. 4, Eq. 5, and Eq. 6 respectively. Finally, the diagnosis performances are the diagnosis error rate (DER) and the diagnosis accuracy rate (DAR), defined by Eq. 7 and Eq. 8.

$$FRR = \frac{FR}{TA + FR} \tag{4}$$

$$FAR = \frac{FA}{FA + TR} \tag{5}$$

$$DetAcc = \frac{TA + TR}{TA + TR + FA + FR} \tag{6}$$

$$DER = \frac{DE}{CD + DE} \tag{7}$$

$$DAR = \frac{CD}{CD + DE} = 1 - DER \tag{8}$$

On the other side, as MDD may be considered a speech recognition task, two scores appear massively in the latest papers, the word error rate (WER), and the phoneme error rate (PER). Given a sequence of N words, the WER is defined as:

$$WER = \frac{S + D + I}{N} \tag{9}$$

where S is the number of substitutions, D is the number of deletions, and I is the number of insertions. Similarly, The PER is the total number of errors made in pronouncing the sequence of words, including substitutions, deletions, and insertions of phonemes. Figure 7 depicts a summary of the performance metrics used in the extracted papers.

The majority of the selected papers used the accuracy and the F1 scores to determine the efficiency of their systems. The DAR is moderately used, although it seems to be the most relevant measure for evaluating this type of task. The top-N error rates were used by [69]. Meanwhile, studies that deal with task clustering, assess the performances mainly using the Davies-Bouldin Index (DBI).

Lately, with the robustness acquired by the DL algorithms over time, the selected papers from the last three years massively use the WER and the PER [24, 27, 29, 31, 32, 36, 38–40, 59, 60, 64].

5 Discussion

In this systematic literature review of studies focusing on the use of deep learning algorithms in mispronunciation detection and diagnosis, we discovered 403 papers out of which 53 were selected as primary studies. We synthesized and analyzed them to obtain responses to the identified research questions. Below, we summarized responses to these questions.

RQ1: The first research question was “What are the different types of papers that were included in this study?”, the question aims to assess whether the subject is topical. The list of selected papers in Table 1 as well as the year-wise distribution in Fig. 2 show a heightened interest in using DL algorithms for MDD. For example, in 2017, the num-

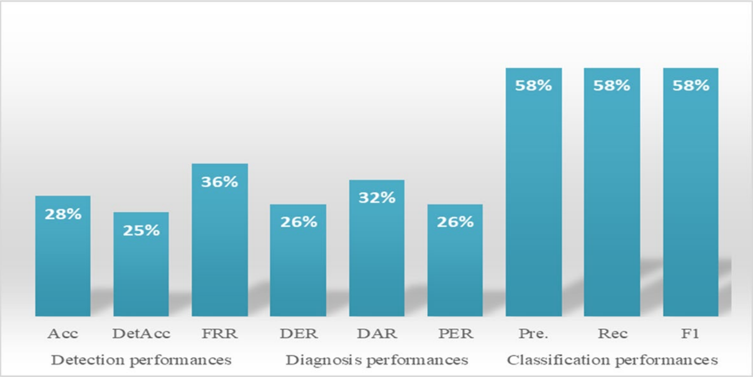

Fig. 7 Summary of the performance metrics used in the selected papers

ber of papers published in journals was three with a total h-index of about 201, while in 2021 the h-index reached 431. This augurs well for the growing reputation of works carried out and their contribution in the various related fields such as signal analysis, pattern recognition, natural language processing, artificial intelligence, etc.

RQ2: The second research question was "What are the different corpora/languages identified in the selected papers?". Traditionally, the challenge in CAPT systems is the lack of speech corpora, with that, the majority of them are private which hinders researchers from training and testing their frameworks. Meanwhile, deep learning algorithms are data-intensive. Among the 53 selected papers, 29 dealt with non-native English corpora. English is the first recommended language to learn to interact with habitants of most countries. Consequently, a wide range of the non-native available corpora has English as the target language (see Table 3). Besides that, China is an emerging nation with millions of projects around the world. Consequently, one of the most important (in terms of variety and duration) corpora used in the CALL context is for learning Chinese Mandarin by English learners (see Table 3). On the other side, private datasets were designed to model L1-L2 phoneme confusion pairs. Indeed, the relation between L1 background and the target language should be considered for the design of nonnative corpora for better feedback; cross-lingual modeling, transfer learning, and self-supervised learning are the trends to handle this issue. In particular, the L1/L2 language abbreviations of Table 8 are listed in Table 7.

RQ3: The third research question was "What are the pronunciation error patterns identified in the selected papers?". As is seen in the results, the prosodic error pronunciation (suprasegmental) presents an open area of research, and a great challenge, principally for tonal languages such as the Chinese Mandarin language, where the five-pitch contour of the language must be taken into consideration.

RQ4: The fourth research question was "What are the different MDD approaches identified in the selected papers?". Early MDD work that incorporated DNNs did so only to replace GMMs, this has led to improved performance in detection and diagnosis as well. DNN-based ASR allows us to overcome the limitation of the ERN approach; DNNs perform free-phone recognition, which can cover all possible mispronunciation patterns. The pronunciation error diagnosis is performed by aligning the canonical phone sequences and the recognized phone sequences. Recently, fully end-to-end DNN models based on raw speech were explored [36] such methods permit the discovery of complex acoustic phenomena and allow their representation, and thus outperform the conventional methods based on hand-crafted acoustic features such as the widely used Mel frequency cepstral coefficients (MFCCs).

RQ5: The fifth research question was "What deep neural network algorithms have been used in the selected papers?". From Table 8, the DNNs have been widely used for the MDD especially for the feature extraction step and for the acoustic modeling. DNNs were used as standalone models, after 2016, they were combined with other models to build hybrid architectures. Afterward, the LSTM model was combined with DNN for the classification of articulatory features [66] and with CNN which was also used for the first time in the MDD [53].

Table 7 Abbreviations of the languages used in Table 8

Arabic	A	European	Eu	Korean	K	Russian	R
Asian	As	German	G	Mandarin	M	Spanish	S
Chinese	C	Italian	I	Mix L1	Mix	Tibetan	T
English	E	Japanese	J	Pakistani	P		

Table 8 Summary of Selected Studies of Deep Learning Algorithms for Mispronunciation Diagnosis

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[22]	2023	Mix/A	LSTM	Class	L2 Arabic	Identification of the articulation problems in Arabic using a text-dependent, and speaker-independent approach. The system improves gender recognition and mispronunciation detection through the use of LSTM networks
[23]	2023	Mix/E	GCN	E2E	L2-Arctic	The use of GCN offers instant feedback on phone-level errors, streamlines dictation, and alignment, and utilizes prior knowledge for accurate phonetic embeddings
[24]	2023	Mix/E	Transfer Learning (TL)	E2E	L2-Arctic	TL improves mispronunciation detection and diagnosis by transferring prior texts, using wav2vec2.0 for robust acoustic representation, and using textual modulation gates and contrastive loss
[25]	2023	Mix/M	Squeezeformer + Bi-LSTM + Transformer	E2E	PSC-Reading	The paper presents a multi-feature, multi-modal MDD method using a squeezeformer encoder, Bi-LSTM network, and Transformer. The model incorporates phoneme information and improves the F1, and diagnostic accuracy on the PSC-Reading Mandarin dataset
[26]	2023	Mix/E	Transformer	E2E	L2-Arctic	Peppanet is a novel neural method based on the transformer model incorporating phonetic, phonological, and acoustic cues
[27]	2023	Mix/E	RNN Transducer (RNN-T)	AM	L2-Arctic	Researchers proposed an autoregressive model of MDD based on a recurrent neural network transducer (RNN-T) at the phone level. The proposed model outperformed the CTC-based methods for L1 Spanish learners in the F1 score and the PER. Moreover, it achieved improvements in the false accept rate by 16.8%

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[28]	2022	Mix/E	CNN/RNN/ BiLSTM/CTC	AM	L2-Arctic	The authors proposed a model of MDD that uses acoustic features, phonetic and linguistic embeddings. The proposed framework extracted the phonetic embeddings from the ASR model trained on large datasets to overcome the data scarcity of the MDD task
[29]	2022	Mix/E	VQ-VAE	AM	L2-Arctic	The paper uses VQ-VAE and Transformer models for training and error mask prediction, respectively. The proposed method employs DL techniques for tasks like error mask prediction and original acoustic unit sequence prediction
[30]	2022	Mix/E	Transformer	E2E	L2-Arctic	The paper uses a discriminative objective function to train E2E MDD models (hybrid CTC-Attention model), aiming to maximize the expected F1 score directly. Data augmentation involves randomly perturbing fractions of phonetic confusing pairs in L2 learners' training utterances, generating artificial pronunciation error patterns
[31]	2022	Mix/A	TL/ TL + CNN-RNN-CTC	AM	KSU speech	TL approach is proposed alongside the fusion of TL and CNN-RNN-CTC. The system detects and diagnoses mispronunciations, recognizes phonemes, and articulatory features, and uses neural text-to-speech technology
[32]	2022	Mix/M	Transformer and (CTC/Attention) hybrid architecture / self-supervised with Wav-2Vec2.0	E2E	PSC-Reading	The paper uses a self-supervised pre-trained speech representation approach to improve MDD performance using Wav2Vec 2.0 and Wavlm models, utilizing CTC/Attention hybrid and Transformer architectures

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[59]	2022	Mix/E	Transformer	E2E	L2-Arctic	The paper presents a non-autoregressive E2E neural modeling approach for English MDD, addressing challenges like slow inference speed and data scarcity
[37]	2021	T/M	CNN-GRU-CTC	E2E	L2 Mandarin	The paper presents a CNN-GRU-CTC AM for mispronunciation detection in Mandarin for Tibetan students. The E2E model eliminates phonemic or graphemic information, achieving better results with an FRR of 7.26%, a DA of 88.35%, and a combined error rate of 14.91%
[35]	2021	S/M	CRNN-CTC: CNN-RNN-CTC	AM	L2 Mandarin	AM of convolutional recurrent neural networks (CRNN) and LSTM combined with connectivity time series classification (CTC) was used to convert acoustic signals into pinyin label sequences. Results were evaluated using initial error rate (IER) and final error rate (CER), and phonetics knowledge was used to identify error-prone points
[60]	2021	M/E	MLP/ LSTM/ GRU/ liGRU	DNN based ASR	EMA-MAE	The paper presents an ASR-based MDD approach. It evaluates the performance using phoneme sequences, human-labeled transcripts, and original prompts. The best-performing system combines acoustic and articulatory features

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[33]	2021	As/A	CNN (AlexNet)	FE	L2 Arabic	The paper introduces two novel speech analysis methods: deep CNN features, a clustering algorithm for the phonemic errors, and an unsupervised phone variation model (PVM) for the prosodic errors. The method achieves 94% accuracy on six mispronounced Arabic pairs and 97% accuracy on an Arabic dataset of 28 individual phones
[64]	2021	Mix/E	CNN—Transformer – CTC	E2E	L2-Arctic	The paper uses the Self-Supervised Pretraining (SSP) model wav2vec2.0 for MDD tasks. It highlights the need for a high-performance MDD for precise diagnoses of pronunciation errors at phonetic and prosodic levels
[34]	2021	Mix/E	TDNN	AM	“RA”	The ASR system developed in this article addresses non-native reading errors in CALL and automated reading tutors, improving targeted interventions and exercises to improve reading skills
[40]	2021	C/E	Transformer	E2E	CU-CHLOE	Two transformer-based models are presented. T-1 is a standard setup with an encoder, a decoder, a projection part, and the Cross-Entropy (CE) loss, and T-2 is based on wav2vec 2.0, a pretraining framework, and includes a CNN feature encoder, Transformer blocks, a projection part, and the Connectionist Temporal Classification (CTC) loss. Both models are trained in an E2E manner and achieve significant improvements over previous models in terms of PER and F-measure

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[36]	2021	Mix/E	SincNet (or CNN)-CTC-ATT	E2E	L2-Arctic	The paper introduces an E2E neural model using a sincnet module for interpretability and performance. Experiments show adaptability and improved accuracy
[38]	2021	J/M	DNN/CNN/GRU	AM	BLCU	A pre-trained approach uses native speech data for non-native mispronunciation verification, addressing data sparsity issues. The model extracts knowledge from unlabeled raw speech, trains with language adversarial training, and incorporates a sinc filter for formant-like feature recognition
[39]	2021	Mix/E	Transformer	E2E	L2-Arctic	The paper introduces a target text-based method for Transformer backbone, improving accuracy and speed, and outperforming the baseline ASR-based model on the L2-Arctic dataset
[13]	2020	J/E	TL DNN	AM	L2 English	The proposed method of non-native AM based on transfer learning relies on the availability of large native speech corpora of the learner’s native language (L1) and target language (L2). This may limit the applicability of the method to language pairs for which such corpora are available
[42]	2020	Mix/E	encoder-(CNN-RNN)-GRU	E2E	L2-Arctic	Integration of multiple stages, such as an acoustic model, a language model, and a Viterbi decoder, into the SED-MDD model. The MDD system achieves 86.35% accuracy on L2-ARCTIC, outperforming existing models like CNN-RNN-CTC

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[41]	2020	Mix/M	Siamese DNN /Siamese CNN	E2E	BLCU	The paper suggests using phone embedding and Siamese networks in a self-supervised approach. Phone segments are projected into acoustic space, and Siamese networks encode feature vectors
[70]	2019	Mix/M	DNN-HMM	AM	Mandarin annotated spoken (MAS) native and non-native	The paper proposes a training approach based on a DNN-HMM model, investigates the benefits of the trained models, and evaluates their performance through experiments as a classification problem
[43]	2019	R/M	TDNN	AM	BLCU	The paper trains fine-grained speech attribute (FSA) models using TDNN on a Chinese corpus and a Russian learner dataset. Experimental results show a 2.2% improvement in mispronunciation detection compared to a segment-based baseline system. The FSA model is theoretically capable of modeling language-universal speech attributes
[44]	2019	C/E	CNN-RNN-CTC	E2E	CU-CHLOE	The paper contributes to the field by introducing the CNN-RNN-CTC model for the MDD problem, which offers improved performance compared to existing approaches. The proposed approach does not require phonemic or graphemic information or forced alignment between linguistic units

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[45]	2019	Eu/M	DNN(baseline)/ BiLSTM	FE tone model-ing Class	iCALL	The paper explores the use of soft-target tone labels and sequential context information for MDD in Mandarin lexical tones for L2 learners with European L1. Three techniques are proposed: extending the tone model to a BiLSTM, characterizing ambiguous pronunciations with soft targets, and extracting segmental tone features. The proposed framework reduces the averaged equal error rate and improves the averaged area under ROC (AUC)
[46]	2019	P/A	CNN(AlexNet)	FE E2E	L2 Arabic	Proposed CNN features-based and TL-based techniques for mispronunciation detection of Arabic phonemes, achieving 92.2% accuracy, and outperforming state-of-the-art techniques
[19]	2019	Mix/E	SDAE	Denoising	Speech Accent Archive + Noise-92	The paper introduces an MDD model using SDAE for noisy English phonemes in mobile learning using rectified linear unit (ReLU) as the activation function
	2019			FE		
[47]	2019	J/M	CNN-LSTM	E2E	BLCU	The paper proposes a CNN-LSTM combination for pronunciation erroneous tendency (PET) detection, combining CNN for robust features and LSTM for time-sensitive PET modeling. Data augmentation techniques are used to reduce data sparsity. The proposed system faces limitations in dealing with acoustic variations and data sparsity

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[20]	2018	C/E	Multi-distribution (MD) -DNN	FE	Supra-CHLOE	The paper uses MD-DNN for automatic lexical stress and pitch accent detection in L2 English speech. The study uses syllable-based prosodic features and binary variables. MD-DNNs achieve high accuracy of 87.9% and 90.2% in lexical stress and pitch accent detection, compared to previous methods like GMMs and prominence models
[65]	2018	C/E	DNN	FE	CU-CHLOE	The paper investigates non-native phonetic patterns (NN-PPs) in L2 speech using an optimized k-means algorithm and DNN. It divides speech into segments and uses cluster sequences to represent native phonetic patterns
[49]	2018	C/E	DNN	AM	CU-CHLOE	The focus of the paper is on the discovery of the L2-extended phoneme set and its application in improving MDD with a DNN model
[48]	2018	C/E	DNN-LSTM	AM	CU-CHLOE	The focus of the paper is on the use of multi-task learning and a feature representation module to improve the performance of the acoustic-phonemic models (APMs) in MDD using a DNN model for the APM
[61]	2018	C/E	MT-DNN	FE	CU-CHLOE	The paper presents three models: Articulatory(A)-APM, R-AAPM, and A-MT-APM, that integrate articulatory features and outperform phoneme-based APMs in precision, recall, and F1-Measure metrics
[67]	2017	G,I/E	DNN-HMM	AM	ISLE	A phonological feature-based pronunciation training system improves mispronunciation detection, diagnosis, and annotation costs through active learning

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[55]	2017	J/E	DNN/ TL(MT-DNN)/ ML-DNN	ArM	L2 English	This paper explores transfer learning-based methods for pronunciation error detection in non-native speech, addressing data limitations and accurate recognition in CAPT systems
[51]	2017	J/M	DNN/ TL (MT-DNN)/ ML-DNN	ArM	BLCU	The paper proposes DNN-based articulatory models for detecting pronunciation errors of L2 learners, using transfer learning and combining task transfer and language transfer learning
[52]	2017	J/M	DNN	ArM	BLCU	The paper presents articulatory models for pronunciation error detection of L2 learners, using deep neural networks and multi-lingual (ML) learning to efficiently train non-native models
[21]	2017	C/E	MD-DNN	Class	Supra-CHLOE	The paper presents an automatic intonation classifier for L2 English speech using MD-DNN. It uses specific labels to represent intonation as rising, upper, lower, or falling. The classifier achieves 93.0% accuracy, detects mispronunciations, and provides diagnostic feedback
[62]	2017	Ch/E	MD-DNN	AM	CU-CHLOE	The paper introduces an acoustic-graphemic-phonemic model (AGPM) using a DNN for MDD in L2 English speech. It outperforms the ERN approach and achieves a lower phone error rate
[66]	2017	Eu/M	DNN-LSTM	FE Class	iCALL	The study uses articulation features, deep neural networks, and LSTM to enhance phonetic mispronunciation detection of L2 learners by combining speech attribute scores with phone information

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[54]	2017	K/E	FFNN	ArM	ETRI	The paper proposes a DNN model for mispronunciation diagnosis using articulatory goodness of pronunciation (AGOP) features, utilizing forced alignment and recognition, and demonstrating effectiveness in F1 score
[50]	2017	J/M	DNN	FE	BLCU	The paper introduces a method using DNN-based log posterior features for labeling candidates, improving consistency, and reducing annotation time. It addresses labeling articulatory-level mispronunciations and provides readable PETs feedback
[53]	2017	J/M	CNN-LSTM	FE	BLCU	The paper proposes a novel approach to detect and provide corrective feedback for erroneous pronunciations, addressing the challenges of context dependency and data sparseness in PET detection
				DL- based ASR		
[57]	2016	J/M	DNN/ TL (MT-DNN)/ ML-DNN	ArM	BLCU	The paper presents DNN-based articulatory models combining target and native speech databases for efficient training and enhanced performance
[68]	2016	Mix/M	DNN	FE	iCALL	The decision tree framework detects phonetic mispronunciations and offers corrective feedback to L2 learners, based on speech attribute features and automatic learning
[63]	2016	Mix/M	DNN	FE Class	iCALL	The paper suggests using neural network classifiers to integrate speech attribute scores and generate segmental pronunciation scores

Table 8 (continued)

Ref	Year	L1 /L2	DL algorithms	Tasks	Nonnative corpus	Contribution
[56]	2016	Eu/M	DNN	AM	iCALL	The paper uses a DNN-based acoustic model for mispronunciation detection, trained on King-ASR-118 mobile speech corpus. The framework incorporates context information and vector space modeling, commonly used in DL models
[58]	2015	J/M	DNN-HMM	AM	BLCU	The paper investigates DNN-based PET detection in CAPT systems, comparing acoustic features and achieving robust detection accuracies
[69]	2015	C/E E/C	DNN-HMM	AM	Isolated Word	An improved approach for mispronunciation detection and diagnosis in L2 learners' speech using DNN-HMM and optimized GOP measure

Abbreviations:AM: Acoustic modeling, ArM: Articulatory modeling, E2E: End-to-end, FE: Feature extraction, Class: Classification

In 2015, Hu, Qian, and Soong [65] were among the first who deal with DL algorithms for MDD purposes. In [65] as in many subsequent works, the DNN was designed to enhance the GOP measure and thus paved the way to DNN-HMM-based MDD. Most of the time, the DNN architecture is that of the MLP with many hidden layers. In [61], the DNN serves to extract the articulatory features that help to model the phonemes' articulation mechanism, and therefore contribute to their accurate representation. Articulatory features were traditionally extracted expensively by electromagnetic articulography (EMA). Meanwhile, multi-distribution DNNs in [20] performed a suprasegmental MDD in terms of lexical stress and pitch accent detection. Herein, the extracted features cover prosodic, lexical, and syntactic ones. The system shows much better accuracy than that achieved with the traditional GMMs. Feature-extractor-dedicated architectures are preferably, autoencoders that allow the dimension reduction in an unsupervised way.

Recently, the CNN-RNN-CTC model was used to implement the E2E systems; this model can detect and diagnose pronunciation errors without the need for handcrafted features. However, it requires a large amount of data for the training. Indeed, the intensive need for speech data (preferably annotated and labeled data) is the main drawback of the DNN models. Some solutions were proposed to tackle this challenge such as transfer learning [24], or audio augmentation techniques [31].

Tables 5, 6, and 8 show that the tendency in DNN models for MDD is CNN as a feature extractor and transformer for DNN-based ASR. In particular, the text-conditioned transformer becomes state-of-the-art to detect pronunciation errors by conditioning the text. However, its performance degrades in the case of languages with complex spelling.

Finally, the lack of explanation is a great drawback inherent to DNN models in case of mispronunciation diagnosis, hence, recently models such as GCN in [23] are proposed to overcome this issue.

RQ6: The sixth and last research question was "What evaluation techniques were used in the selected papers?", herein, the results in Fig. 7 underline the need for more relevant measures such as DAR and DER. Meanwhile, the PER is established as the de facto measure to assess MDD systems, as those systems are close to ASR ones in an E2E way.

6 Conclusion

We identified 403 articles from searching the literature, of which 53 papers published between the years 2015 and 2023 were selected as primary studies. Our work presented the results gathered after performing a systematic literature review to identify and evaluate the current approaches in using deep learning algorithms for mispronunciation detection and diagnosis. The results of this review show that MDD is a highly active area of research.

In the majority of the papers, DNN models have been mainly used in the acoustic modeling of speech instead of GMMs, when estimating observation probabilities in the states of HMMs. Moreover, the utilized corpora in the included study were mostly public in English or Mandarin, and the environments were mostly non-noisy.

It is surprising to see that most of the papers still use MFCCs as feature extraction for speech signals in deep learning models. However, recent systems used DNNs as feature extractors, particularly the CNN algorithm. Another observation is that there is little work on the use of DNNs as generative models.

While the DNN models offer an alternative to the early ERNs which required expert knowledge, they need a huge amount of data for their training. The use of transfer learning throughout CNN models such as AlexNet may mitigate this need.

On the other side, DNN models may not detect subtle mispronunciations and struggle to detect insertions that were handled by the ERNs.

Finally, the results show that there is no complete study with satisfactory outcomes. It is necessary to test more techniques of deep learning on different types of features and different metrics for evaluation.

Data availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the article's content.

Conflict of interests "Not Applicable".

References

1. Shahin M, Ahmed B (2019) Anomaly detection based pronunciation verification approach using speech attribute features. Speech Commun 111:29–43. https://doi.org/10.1016/j.specom.2019.06.003
2. Cohen M, Murveit H, Bernstein J, Price P, Weintraub M (1990) The decipher speech recognition system. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Albuquerque, pp 77–80. https://doi.org/10.1109/ICASSP.1990.115541
3. Eskenazi M (2009) An overview of spoken language technology for education. Speech Commun 51(10):832–844. https://doi.org/10.1016/j.specom.2009.04.005
4. Chen NF, 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). IEEE, Jeju, pp 1–7. https://doi.org/10.1109/APSIPA.2016.7820782
5. Franco H, Neumeyer L, Kim Y, Ronen O (1997) Automatic pronunciation scoring for language instruction. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, Munich, pp 1471–1474. https://doi.org/10.1109/ICASSP.1997.596227
6. Witt SM, Young SJ (2000) Phone-level pronunciation scoring and assessment for interactive language learning. Speech Commun 30(2–3):95–108. https://doi.org/10.1016/S0167-6393(99)00044-8
7. Bahi H, Necibi K (2020) Fuzzy logic applied for pronunciation assessment. Int J Comput Assisted Lang Learn Teach 10(1):60–72. https://doi.org/10.4018/IJCALLT.2020010105
8. Neumeyer L, Franco H, Digalakis V, Weintraub M (2000) Automatic scoring of pronunciation quality. Speech Commun 30(2–3):83–93. https://doi.org/10.1016/S0167-6393(99)00046-1
9. Strik H, Truong KP, Wet FD, Cucchiarini C (2007) Comparing classifiers for pronunciation error detection. 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium, pp 1837–1840. https://doi.org/10.21437/interspeech.2007-512
10. Harrison AM, Lo WK, Qian XJ, Meng H (2009) Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In: International Workshop on Speech and Language Technology in Education (SLaTE), Warwickshire, pp 45–48
11. Wang YB, Lee LS (2015) Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning. IEEE ACM Trans Audio Speech Lang Process 23(3):564–579. https://doi.org/10.1109/taslp.2014.2387413
12. Lee A, Chen NF, Glass J (2016) Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016. IEEE, 6145–6149. https://doi.org/10.1109/icassp.2016.7472858

13. 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 Trans Audio Speech Lang Process 28:391–401. https://doi.org/10.1109/taslp.2019.2955858
14. Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007–001. Keele University and Durham University
15. Neri A, Cucchiarini C, Strik H, Boves L (2002) The pedagogy-technology interface in computer assisted pronunciation training. Comput Assisted Lang Learn 15(5):441–467. https://doi.org/10.1076/call.15.5.441.13473
16. Witt SM (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 (ISADEPT), Stockholm, pp 1–8
17. Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Mag 29(6):82–97. https://doi.org/10.1109/msp.2012.2205597
18. Agarwal C, Chakraborty P (2019) A review of tools and techniques for computer aided pronunciation training (CAPT) in English. Educ Inf Technol 24(6):3731–3743. https://doi.org/10.1007/s10639-019-09955-7
19. Wu Y, Zhang J, Dong Q (2019) The use of SDAE in noisy English mispronunciation detection and diagnosis towards application in mobile learning. In: International Symposium on Signal Processing Systems (SSPS). ACM, Beijing, pp 176–180. https://doi.org/10.1145/3364908.3365302
20. Li K, Mao S, Li X, Wu Z, Meng H (2018) Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. Speech Commun 96:28–36. https://doi.org/10.1016/j.specom.2017.11.003
21. Li K, Wu X, Meng H (2017) Intonation classification for L2 English speech using multi-distribution deep neural networks. Comput Speech Lang 43:18–33. https://doi.org/10.1016/j.csl.2016.11.006
22. Ahmed A, Bader M, Shahin I, Nassif AB, Werghi N, Basel M (2023) Arabic Mispronunciation Recognition System Using LSTM Network. Information 14(7):413. https://doi.org/10.3390/info14070413
23. Yan BC, Wang HW, Wang YC, Chen B (2023) Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, IEEE, pp 1–5. https://doi.org/10.1109/icassp49357.2023.10097226
24. Peng L, Gao Y, Bao R, Li Y, Zhang J (2023) End-to-End Mispronunciation Detection and Diagnosis Using Transfer Learning. Appl Sci 13(11):6793. https://doi.org/10.3390/app13116793
25. Guo S, Kadeer Z, Wumaier A, Wang L, Fan C (2023) Multi-Feature and Multi-Modal Mispronunciation Detection and Diagnosis Method Based on the Squeezeformer Encoder. IEEE Access 11:66245–66256. https://doi.org/10.1109/access.2023.3278837
26. Yan BC, Wang HW, Chen B (2023) Peppanet: Effective mispronunciation detection and diagnosis leveraging phonetic, phonological, and acoustic cues. In: Spoken Language Technology Workshop (SLT). IEEE, Doha, pp 1045–1051. https://doi.org/10.1109/slt54892.2023.10022472
27. Zhang DY, Saha S, Campbell S (2023) Phonetic RNN-transducer for mispronunciation diagnosis. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Rhodes Island, pp 1–5. https://doi.org/10.1109/icassp49357.2023.10094945
28. 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: International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, IEEE, pp 6827–6831. https://doi.org/10.1109/icassp43922.2022.9746604
29. Zhang Z, Wang Y, Yang J (2022) Masked acoustic unit for mispronunciation detection and correction. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Singapore, pp 6832–6836. https://doi.org/10.1109/icassp43922.2022.9747414
30. Yan BC, Wang HW, Jiang SW, Chao FA, Chen B (2022) Maximum f1-score training for end-to-end mispronunciation detection and diagnosis of L2 English speech. In: International Conference on Multimedia and Expo (ICME). IEEE, Taipei, pp 1–5. https://doi.org/10.1109/icme52920.2022.9858931
31. Algabri M, Mathkour H, Alsulaiman M, Bencherif MA (2022) Mispronunciation detection and diagnosis with articulatory-level feedback generation for non-native arabic speech. Mathematics 10(15):2727. https://doi.org/10.3390/math10152727
32. Shen Y, Liu Q, Fan Z, Liu J, Wumaier A (2022) Self-Supervised Pre-Trained Speech Representation Based End-to-End Mispronunciation Detection and Diagnosis of Mandarin. IEEE Access 10:106451–106462. https://doi.org/10.1109/access.2022.3212417

33. Nazir F, Majeed MN, Ghazanfar MA, Maqsood M (2021) A computer-aided speech analytics approach for pronunciation feedback using deep feature clustering. Multimed Syst 29(3):1699–1715. https://doi.org/10.1007/s00530-021-00822-5
34. Qin Y, Qian Y, Loukina A, Lange P, Misra A, Evanini K, Lee T (2021) Automatic detection of word-level reading errors in nonnative English speech based on ASR output. In: International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, Hong Kong, pp 1–5. https://doi.org/10.1109/iscslp49672.2021.9362102
35. Huang Y (1952) Huang Y (2021) Detection of Mispronunciation in Non-native Speech Using Acoustic Model and Convolutional Recurrent Neural Networks. J Phys Conf Ser 3:032043. https://doi.org/10.1088/1742-6596/1952/3/032043
36. Yan BC, Chen B (2021) End-to-end mispronunciation detection and diagnosis from raw waveforms. In: European Signal Processing Conference (EUSIPCO). IEEE, Dublin, pp 61–65. https://doi.org/10.23919/eusipco54536.2021.9615987
37. Gan Z, Zhao X, Zhou S, Wang R (2021) Improving mispronunciation detection of Mandarin for Tibetan students based on the end-to-end speech recognition model. In: International Symposium on Artificial Intelligence and its Application on Media (ISAIAM). IEEE, Xi'an, pp 151–154. https://doi.org/10.1109/isaiam53259.2021.00039
38. Yang L, Fu K, Zhang J, Shinozaki T (2021) Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning. Neural Netw 142:597–607. https://doi.org/10.1016/j.neunet.2021.07.017
39. Zhang Z, Wang Y, Yang J (2021) Text-conditioned transformer for automatic pronunciation error detection. Speech Commun 130:55–63. https://doi.org/10.1016/j.specom.2021.04.004
40. Wu M, Li K, Leung WK, Meng H (2021) Transformer based end-to-end mispronunciation detection and diagnosis. Interspeech, ISCA, Brno, pp 3954–3958. https://doi.org/10.21437/interspeech.2021-1467
41. Xie Y, Wang Z, Fu K (2020) L2 Mispronunciation Verification Based on Acoustic Phone Embedding and Siamese Networks. J Signal Process Syst. https://doi.org/10.1007/s11265-020-01598-z
42. Feng Y, Fu G, Chen Q, Chen K (2020) SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, IEEE, pp 3492–3496. https://doi.org/10.1109/icassp40776.2020.9052975
43. Guo M, Rui C, Wang W, Lin B, Zhang J, Xie Y (2019) A study on mispronunciation detection based on fine-grained speech attribute. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, Lanzhou, pp 1197–1201. https://doi.org/10.1109/APSIPAASC47483.2019.9023156
44. Leung WK, Liu X, Meng H (2019) CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Brighton, pp 8132–8136. https://doi.org/10.1109/ICASSP.2019.8682654
45. Li W, Chen NF, Siniscalchi SM, Lee CH (2019) Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and BLSTM-based deep tone models. IEEE ACM Trans Audio Speech Lang Process 27(12):2012–2024. https://doi.org/10.1109/TASLP.2019.2936755
46. Nazir F, Majeed MN, Ghazanfar MA, Maqsood M (2019) Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for Arabic phonemes. IEEE Access 7:52589–52608. https://doi.org/10.1109/ACCESS.2019.2912648
47. Yang L, Xie Y, Zhang J (2019) Pronunciation Erroneous Tendency Detection with Combination of Convolutional Neural Network and Long Short-Term Memory. Int J Asian Lang Process 28(2):49–66
48. Mao S, Wu Z, Li R, Li X, Meng H, Cai L (2018) Applying multitask learning to acoustic-phonemic model for mispronunciation detection and diagnosis in l2 English speech. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, IEEE, pp 6254–6258. https://doi.org/10.1109/ICASSP.2018.8461841
49. Mao S, Li X, Li K, Wu Z, Liu X, Meng H (2018) Unsupervised discovery of an extended phoneme set in l2 English speech for mispronunciation detection and diagnosis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol 2018. IEEE, Calgary, pp 6244–6248. https://doi.org/10.1109/ICASSP.2018.8462635
50. Wei X, Chen J, Wang W, Xie Y, Zhang J (2017) A study of automatic annotation of PETs with articulatory features. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, Kuala Lumpur, pp 1608–1612. https://doi.org/10.1109/APSIPA.2017.8282281

51. 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 TRANS Inf Syst E100.D(9):2174–2182. https://doi.org/10.1587/transinf.2017edp7019
52. Duan R, Kawahara T, Dantsuji M, Zhang J (2017) Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, New Orleans, pp 5815–5819. https://doi.org/10.1109/ICASSP.2017.7953271
53. Yang L, Xie Y, Gao Y, Zhang J (2017) Improving pronunciation erroneous tendency detection with convolutional long short-term memory. In: International Conference on Asian Language Processing (IALP). IEEE, Singapore, pp 52–56. https://doi.org/10.1109/IALP.2017.8300544
54. Ryu H, Chung M (2017) Mispronunciation diagnosis of L2 English at articulatory level using articulatory goodness-of-pronunciation features. In: Workshop on Speech and Language Technology in Education (SLaTE). ISCA, Stockholm, pp 65–70. https://doi.org/10.21437/slate.2017-12
55. Duan R, Kawahara T, Dantsuji M, Nanjo H (2017) Transfer learning based non-native acoustic modeling for pronunciation error detection. In: 7th ISCA Workshop on Speech and Language Technology in Education (SLaTE). ISCA, Stockholm, pp 42–46. https://doi.org/10.21437/slate.2017-8
56. Tong R, Chen NF, Ma B, Li H (2016) Context aware mispronunciation detection for mandarin pronunciation training. Interspeech, San Francisco, ISCA, pp 3112–3116. https://doi.org/10.21437/interspeech.2016-289
57. Duan R, Kawahara T, Dantsuji M, Zhang J (2016) Multi-lingual and multi-task DNN learning for articulatory error detection. In: Asia- Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE, Jeju, pp 1–4. https://doi.org/10.1109/APSIPA.2016.7820800
58. Gao Y, Xie Y, Cao W, Zhang J (2015) A study on robust detection of pronunciation erroneous tendency based on deep neural network. Interspeech, Dresden, ISCA, pp 693–696. https://doi.org/10.21437/interspeech.2015-242
59. Wang HW, Yan BC, Chiu HS, Hsu YC, Chen B (2022) Exploring non-autoregressive end-to-end neural modeling for English mispronunciation detection and diagnosis. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Singapore, pp 6817–6821. https://doi.org/10.1109/ICASSP43922.2022.9747569
60. Khanal S, Johnson MT, Soleymanpour M, Bozorg N (2021) Mispronunciation detection and diagnosis for Mandarin accented English speech. In: International Conference on Speech Technology and Human-Computer Dialogue (SpeD). IEEE, Bucharest, pp 62–67. https://doi.org/10.1109/SpeD53181.2021.9587408
61. Mao S, Wu Z, Li X, Li R, Wu X, Meng H (2018) Integrating articulatory features into acoustic phonemic model for mispronunciation detection and diagnosis in l2 English speech. In: International Conference on Multimedia and Expo (ICME). IEEE, San Diego, pp 1–6. https://doi.org/10.1109/ICME.2018.8486462
62. Li K, Qian X, Meng H (2017) Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. IEEE ACM Trans Audio Speech Lang Process 25(1):193–207. https://doi.org/10.1109/TASLP.2016.2621675
63. Li W, Siniscalchi SM, Chen NF, Lee CH (2016) Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. In: International conference on acoustics, speech and signal processing (ICASSP). IEEE, Shanghai, pp 6135–6139. https://doi.org/10.1109/ICASSP.2016.7472856
64. Peng L, Fu K, Lin B, Ke D, Zhang J (2021) A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis. Interspeech, Brno, ISCA, pp 4448–4452. https://doi.org/10.21437/interspeech.2021-1344
65. Li X, Mao S, Wu X, Li K, Liu X, Meng H (2018) Unsupervised discovery of non-native phonetic patterns in L2 English speech for mispronunciation detection and diagnosis. Interspeech, Hyderabad, ISCA, pp 2554–2558. https://doi.org/10.21437/interspeech.2018-2027
66. Li W, Chen NF, Siniscalchi SM, Lee CH (2017) Improving mispronunciation detection for nonnative learners with multisource information and LSTM-based deep models. In: Interspeech. ISCA, Stockholm, pp 2759–2763. https://doi.org/10.21437/interspeech.2017-464
67. Arora V, Lahiri A, Reetz H (2017) Phonological feature based mispronunciation detection and diagnosis using multi-task DNNs and active learning. Interspeech, Stockholm, ISCA, pp 1432–1436. https://doi.org/10.21437/interspeech.2017-1350
68. Li W, Li K, Siniscalchi SM, Chen NF, Lee CH (2016) Detecting mispronunciations of L2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees. In: Interspeech. ISCA, San Francisco, pp 3127–3131. https://doi.org/10.21437/interspeech.2016-517

69. Hu W, Qian Y, Soong FK (2015) An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech. In: Workshop on Speech and Language Technology in Education (SLaTE). ISCA, Leipzig, pp 71–76
70 Chen B, Hsu YC (2019) Mandarin Chinese mispronunciation detection and diagnosis leveraging deep neural network based acoustic modeling and training techniques. In: Lu X, Chen B (eds) Computational and Corpus Approaches to Chinese Language Learning. Chinese Language Learning Sciences, Springer, Singapore, pp 217–234. https://doi.org/10.1007/978-981-13-3570-9_11
71. Raux A, Kawahara T (2002) Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer assisted pronunciation learning. In: International Conference on Spoken Language Processing (ICSLP). ISCA, Denver, pp 737–740. https://doi.org/10.21437/icslp.2002-241
72. Cheng J, Chen X, Metallinou A (2015) Deep neural network acoustic models for spoken assessment applications. Speech Commun 73:14–27. https://doi.org/10.1016/j.specom.2015.07.006
73. Jiang SW, Yan BC, Lo TH, Chao FA, Chen B (2021) Towards robust mispronunciation detection and diagnosis for L2 English learners with accent-modulating methods. In: Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, Cartagena, pp 1065–1070. https://doi.org/10.1109/ASRU51503.2021.9688291
74. Kim S, Gholami A, Shaw A, Lee N, Mangalam K, Malik J, Mahoney MW, Keutzer K (2022) Squeezeformer: An efficient transformer for automatic speech recognition. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, pp 9361–9373
75. Qian X, Meng H, Soong F (2010) Capturing L2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (CAPT). In: International Symposium on Chinese Spoken Language Processing, IEEE, Tainan, pp 84–88. https://doi.org/10.1109/iscslp.2010.5684845

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.