Journal of Infrastructure, Policy and Development 2025, 9(2), 10078.
https://doi.org/10.24294/jipd10078

Review

The architecture of automatic scoring systems for non-native English spontaneous speech: A systematic literature review

Un I. Kuok

Institute of Collaborative Innovation, University of Macau, Macao SAR 999078, China; kuokuni.translator@outlook.com

CITATION

Kuok UI. (2025). The architecture of automatic scoring systems for non-native English spontaneous speech: A systematic literature review. Journal of Infrastructure, Policy and Development. 9(2): 10078. https://doi.org/10.24294/jipd10078

ARTICLE INFO

Received: 4 November 2024
Accepted: 4 December 2024
Available online: 7 April 2025

Abstract: Given the heavy workload faced by teachers, automatic speaking scoring systems provide essential support. This study aims to consolidate technological configurations of automatic scoring systems for spontaneous L2 English, drawing from literature published between 2014 and 2024. The focus will be on the architecture of the automatic speech recognition model and the scoring model, as well as on features used to evaluate phonological competence, linguistic proficiency, and task completion. By synthesizing these elements, the study seeks to identify potential research areas, as well as provide a foundation for future research and practical applications in software engineering.

Keywords: automatic scoring system; automatic speech recognition; L2 English speaking; spontaneous speech; assessment and evaluation

1. Introduction

In any educational systems, faculty often face substantial workloads, including extensive testing and grading responsibilities. This highlights the need for accessible training tools that enable students to practice and improve their skills independently. To release faculty members from these assessment tasks and allow them to re-allocate their time to fulfil other obligations, such as course design, as well as provide students access to an independent English speaking training tool, the development of automatic scoring systems has been gaining traction. Among these efforts are those under the research umbrella of Computer-Assisted Language Learning (CALL).

While the automatic scoring system for English writing has been well-researched, and commercial applications, such as Grammarly, well-established, the demanding technological requirements for automated English-speaking assessment, especially for spontaneous or unrestricted speech, which is also more data demanding in training process than read speech, has slowed the pace for research in this area (Cheng et al., 2015). Nevertheless, recent technical advancement with machine learning, neural networks, and transformer-based learning models have broken this bottleneck. Although systematic literature reviews exist summarizing the technical specifications of automatic speech assessment, none specifically address the niche of non-native, spontaneous English speech. To bridge this gap, the present study aims to examine the techniques employed in the automatic scoring of spontaneous L2 (second language) English speaking tasks.

2. Literature review

In the following sections, existing literatures on the construction of automatic speech recognition systems, as well as their application in recognizing non-native and spontaneous speeches will be summarized.

2.1. Automatic speech recognition (ASR) systems

A typical automatic speech scoring system consists of a speech recognizer and scoring model (**Figure 1**). Based on the audio input, the speech recognizer generates signals through which speech features are extracted. Subsequently, these features are fitted into the scoring model for grading purposes.

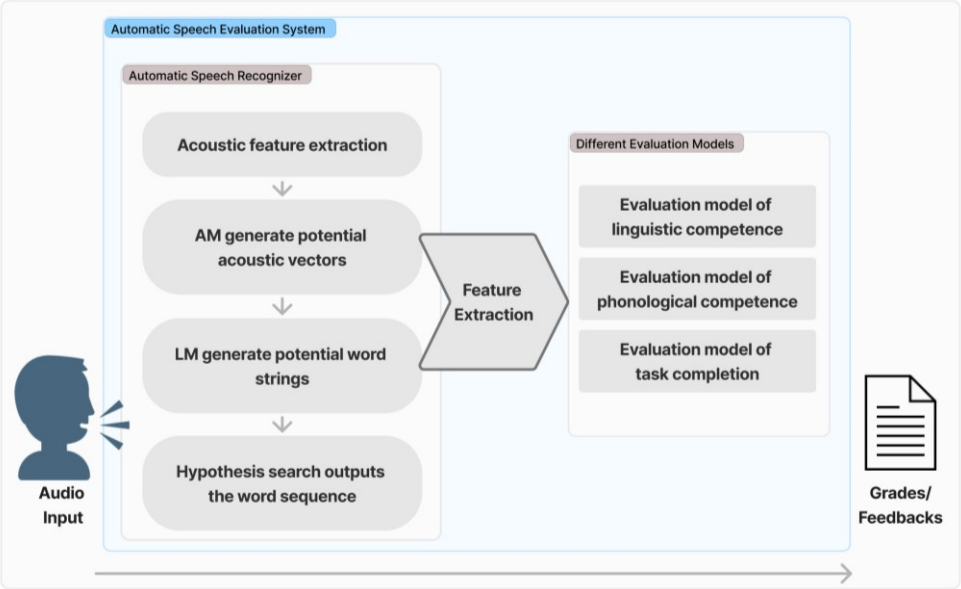

Figure 1. A simplified architecture for an automatic speech evaluation system.

Being the fundament of automatic scoring system, an automatic speech recognition system normally has four main components: 1. signal processing and feature extraction, 2. acoustic model (AM) relevant with the pronunciation, 3. language model (LM) related to the language patterning, and 4. hypothesis search (Saon and Chien, 2012; Yu and Deng, 2015).

During the recognition process, the audio input is pre-processed to remove the background noise and correct distorted channels before it undergoes framing and windowing processes to divide the input into smaller segments with minimal edging effects (Lee et al., 2022; Tamazin et al., 2019). These segments are then converted into time and frequency domains and salient feature vectors applicable to the AM and LM are extracted (Alharbi et al., 2021; Saon and Chien, 2012; Yu and Deng, 2015), using common feature extraction techniques, including MFCC (Mel Frequency Cepstral Coefficients), LDA (Linear Discriminant Analysis) and Probabilistic LDA (Fendji et al., 2022). The AM and LM then calculate the probability of potential word sequences, and the hypothesis search outputs the word sequence with the highest combined AM and LM probability.

2.1.1. Acoustic modeling in ASR models

Taking in the feature vectors, the AM will generate the possible acoustic vectors, or statistical representations for the distinct sounds, for which the AM scores will be calculated referencing to the acoustics and phonetics knowledge in the database (Benkerzaz et al., 2019; Fendji et al., 2022; Saon and Chien, 2012; Yu and Deng, 2015).

The most prevalent acoustic modelling approach is the Gaussian mixture model—Hidden Markov Model (GMM-HMM), or HMM. In this approach, acoustic signals extracted using MFCC are modelled into probability distributions (Saon and Chien, 2012). HMMs consist of a hidden part and an observable part (Trentin and Gori, 2001). The hidden states, which are abstract and non-observable, conceptualize the phonemes in speech—the observable events in the real world; while the observable states are the statistical representations of these observable events (Trentin and Gori, 2001). Forward backward (or Baum Welch) and Viterbi are two popular algorithms applied in HMM (Trentin and Gori, 2001), with both of them based on the general maximum-likelihood (ML) criterion.

Other more advanced acoustic modeling approaches include neural network-based acoustic models and end-to-end models. Unlike traditional GMM-HMM models, which require users to train a feature selection model along with the AM and LM, neural network-based models can learn the required features, such as fluency features, directly from the raw data (Liu et al., 2023; Sainath et al., 2017; Yu and Deng, 2015). Moreover, the recognition capability of DNN-HMM acoustic models has outperformed that of GMM-HMM models, whose performance improvement becomes saturated after a certain number of hours of training data (Cheng et al., 2015). On the other hand, end-to-end ASRs streamline the training process by using a single criterion, eliminating the need to train the AMs, LMs, and feature selection models separately—A process that discounts recognition performance (Miao and Metze, 2017).

2.1.2. Language modeling in ASR models

Subsequently, the language model analyses the statistical representation of the acoustic vectors referencing to its corpora and proposes the most likely word sequences (Fendji et al., 2022). There are two common types of LM, namely grammar-based (or deterministic LM) and statistic-based (or stochastic LM). Grammar-based LMs, designed by linguists to construct the grammatical framework, specify the possible word sequences referencing to collocation and grammatical rules in the database (Rosenfeld, 2000). On the other hand, statistic-based LMs is represented by P(W), the probabilities of the word string w_i, taking into consideration the previous word strings $h_i = w_1, \ldots, w_{i-1}$ (Fendji et al., 2022):

$$\mathrm{P}(W) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \ldots, w_{i-1})$$

$$= \prod_{i=1}^{n} P(w_i | h_i)$$

However, the complexity of this LM accumulates as time go by during the recognition. To mitigate this issue, the N-gram approach based on word classes (e.g. nouns, verbs, adjectives and adverbs) is commonly applied where the probability P(W) of $w_I$is only determined by the probability of the previous n-1 word string(s) (Brown et al., 1992; Fendji et al., 2022; Saon and Chien, 2012):

$$\mathrm{P}(W) = \prod_{i=1}^{n} P(w_i | w_{i-n+1,} \ldots, w_{i-1})$$

2.2. ASR for non-native utterance

The automated recognition of non-native speakers' spontaneous speech is a challenging task, as evidenced by the error rate of speech recognizers developed for this task. For instance, Chen and Zechner (2011) reported a 30% word error rate (WER) in speech recognition, and these frequent errors at the recognition stage negatively affect the subsequent stages of the speech scoring system in general.

L2 speakers often carry the traits of their native accent into their second language (Georgakis et al., 2016). When these traits are analyzed as acoustic signals in an ASR system, they manifest as differences in the energy of phoneme classes, formant features, and frequency (Arslan and Hansen, 1996; Liu and Fung, 1999), as well as their pause patterns (Cucchiarini et al., 2000) compared to those of native speakers

The challenges posed by the fundamental differences between native and non-native utterances are significant for ASR models. Empirical studies have consistently demonstrated that recognition accuracy for accented speech is lower than that for native speech (Wills et al., 2023; Yu and Deng, 2015). While the performance of ASR systems for non-native utterances can be improved by using specialized models tailored to different accents, there has also been considerable research aimed at developing systems capable of identifying various accents through their acoustic features (Arslan and Hansen, 1996; Ge, 2015; Liu and Fung, 1999). Additionally, training language models with common non-native pronunciation patterns has shown promise (Livescu and Glass, 2000). More recent studies have explored the construction of accent-independent speech recognition systems using advanced machine learning algorithms. These include pre-training wav2vec 2.0 with unsupervised datasets (Aksënova et al., 2022), applying standard time-delayed neural network structure (TDNN) pre-trained with the Kaldi toolkit (Li et al., 2021) and maximum likelihood linear regression (MLLR) algorithm (Deng et al., 2007).

2.3. ASR for spontaneous speeches

Spontaneous speech is commonly believed to be characterized by unclear pronunciation and disfluency, such as fillers or filled pauses (um, uh) and corrections (Lease et al., 2006; Van Bergem, 1995). Furthermore, the grammatical mistakes in spontaneous utterances will increase the number of errors in the recognition (Knill et al., 2019).

When considering unclear pronunciation, interpreted acoustically, spontaneous speech has a more concentrated and limited spectral profile, which implies a relatively constrained pitch range (Nakamura et al., 2008), its phonemes are shorter and scattered

in the acoustic space (Gerosa et al., 2006; Shriberg, 1999), as well as more varied and less uniform than those in read speech, indicating that phonemes are pronounced with greater diversity (Nakamura et al., 2008). Since most AM in ASR models are trained with read speech corpora, the different acoustic features between read and spontaneous speech limit the AM's adaptability, making ASR models to often underperform in spontaneous speech (Gabler et al., 2023; Yu and Deng, 2015).

Alongside unclear pronunciations, ASR models also need to overcome disfluencies in spontaneous speech. A typical challenge in recognising spontaneous speech is the prevalence of self-correction, which involves disfluencies such as false starts and repetition (Dufour et al., 2014; Lease et al., 2006). These phenomena often exhibit cross-serial dependencies, which are complex and difficult to model using traditional grammar-based LMs which incorporate context-free or finite-state grammars as foundational structures; as well as for statistical language models that relay on local context (N-gram) rather than these long-range dependencies (Heeman and Allen, 1999; Lease et al., 2006). Also, the inaccurate annotation of disfluency tags—such as editing terms, repetitions, repairs, and false starts—in the ASR models will affect the automatic grading system (Yoon and Bhat, 2018).

To improve the recognition accuracy of the spontaneous speech, Heeman and Allen (1999) proposed a language model employing PoS (part of speech) tagging and discourse marker identification to identify the segmenting turns or interruption points. Other approaches include the TAG-based model of speech repairs that identifies the reparanda within an utterance (Johnson and Charniak, 2004). Building on this model, Lease et al. (2006) proposed a TAG-based model of speech repairs with a maximum-entropy ranker that predicts not only the reparanda but also the fillers.

2.4. Automated assessment for oral language proficiency

Most of the automated assessment for oral language skills were error-based, using ASR to identify various types of error in the students' utterance, such as grammatical errors (Knill et al., 2019; Yoon and Bhat, 2018), pronunciation errors (Chen et al., 2019). In addition to error-base studies, researchers also developed systems to assess the fluency of speech (Cucchiarini et al., 2002) and prosody of the delivery. While some studies focused on read aloud utterance (Cucchiarini et al., 2000; Molenaar et al., 2023), others examined conversational response and spontaneous speech on a given topic (Cucchiarini et al., 2002; Zechner et al., 2014).

Various features extracted by ASR models have been studied for agreement with human raters (Cucchiarini et al., 2000; Kobayashi and Abe, 2016; Zechner et al., 2014). Correlation studies have identified that speech rate, articulation rate, phonation-time ratio, number of silent pauses, total pause duration, and mean length of speech runs are acoustic features that correlate with human raters' assessments of perceived fluency (Cucchiarini et al., 2000). Additionally, linguistic variables such as tokens, types, and nouns have been found to correlate with human graders' evaluations of students' oral English proficiency (Kobayashi and Abe, 2016).

Aiming to identify research gaps for future studies, the present study systematically reviews the methodologies applied in the modelling of the automatic

scoring system for non-native English spontaneous speech in existing literature, specifically through the following research questions:

1) What methods or algorithms are used to model the ASR and scoring components in automatic scoring systems?
2) What features are extracted through automatic scoring systems for scoring.

3. Methodology

To systematically review the methodological components incorporated in these models and identify areas for further investigation, the current study will adopt the PRISMA framework to review the research regarding automatic English scoring system for spontaneous speech over the past decade (2014–2024).

3.1. Bibliometrics and tool selection

A systematic search was performed using four databases: Web of Science, ERIC, IEEE and Scopus, all of them last consulted on 1 May, 2024. The search syntax for the databases was:

("automatic" OR "automated" OR "automation" OR "machine learning" OR "artificial intelligence" OR "AI") AND ("assessment" OR "assess" OR "score" OR "scoring" OR "grade" OR "grading" OR "evaluate" OR "evaluation" OR "test" OR "exam") AND ("English speaking" OR "spoken English" OR "oral English") AND ("spontaneous speech" OR "dialog" OR "dialogue" OR "conversation")

After the search, a total of 139 studies were found (80 in Web of Science, 30 in Scopus, 21 in IEEE Xplore, 8 in ERIC). An initial check of duplication removed 21 records from the screening. All the remaining research was examined according to the following criteria **(Figure 2)**:

1) The study should be automatically assessing L2 English speaking tasks.
2) The study should be automatically assessing spontaneous speech.
3) The study should include the details of the automatic scoring system, namely the ASR and the scoring model.

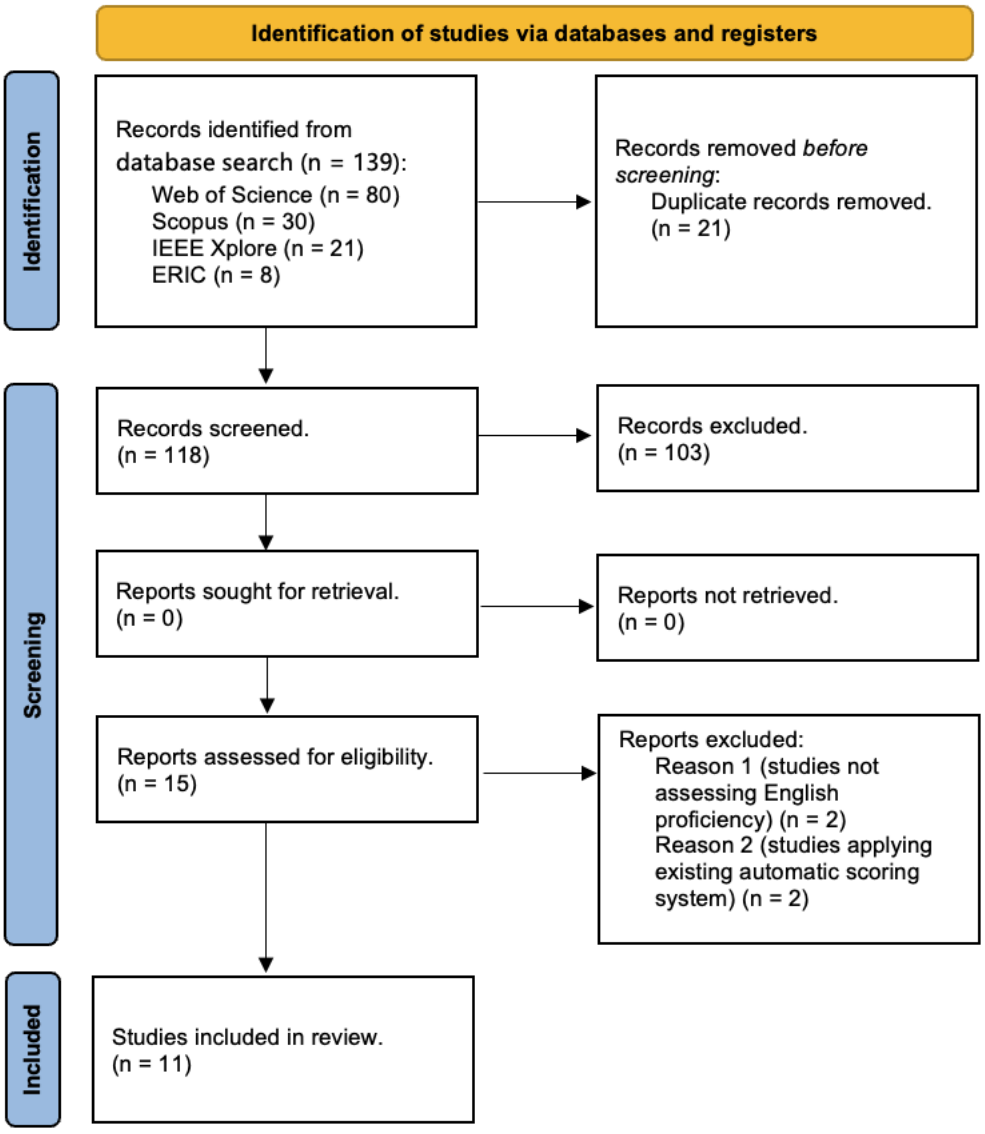

Figure 2. Flow diagram of research article selection process (Page et al., 2021).

3.2. Coding procedures

To address the research questions, information relevant to the predefined categories was extracted from the 11 studies included in the review. The focus of the study, the components of the ASR models, the features extracted for scoring, and the scoring model were consolidated in **Table 1**.

Table 1. Summary of the construction of automatic scoring systems.

	Author(s)	Focus	Components of the ASR	Features extracted for scoring	Scoring model
1	(Bhat and Yoon, 2015)	Syntactic complexity	**AM:** HMM recognizer; Gender independent triphone acoustic model; **LM:** Bigram LM Trigram LM Four-gram LM **Feature selection:** Vector-space model (VSM)	**Syntactic complexity is represented by POS tag** **Model 1. POS-based vector space model:** cos_i: cosine similarity value of the test response with the representative vector of score level i = 1,2,3,4; cosmax: the score level with the highest similarity score given the response. **Model 2. POS language models:** lm_i: logprob (likelihood) of the LM of score level i = 1, 2, 3, 4; lmmax: the score level of the LM with the maximum logprob given the response.	Multiple linear regression
		Fluency		HMM acoustic model score (amscore) speaking rate (wpsec)types per second (tpsecutt) average chunk length in words (wdpchk) global normalized language model score (lmscore)	SpeechRater
2	(Qian et al., 2016)	Task completion	**AM:** GMM-HMM DNN-HMM **LM:** Two trigram LMs **Feature extraction and transformation:** MFCCs LDA and MLLT fMLLR i-Vector Extraction	Latent semantic analysis Content vector analysis Confidence score	Random forest regressor

Table 1. (*Continued*).

	Author(s)	Focus	Components of the ASR	Features extracted for scoring	Scoring model
3	(Tao et al., 2016)	Fluency; Linguistic proficiency	**AM:** GMM-HMM DMM-HMM Tandem **LM:** tri-gram language model	Fluency number of words per second number of words per chunk number of silences average duration of silences frequency of long pauses (≥ 0.5 sec.) number of filled pauses (uh and um) Frequency of between-clause silences edit disfluencies compared to within-clause silences edit disfluencies Rhythm, Intonation & Stress overall percentages of prosodic events mean distance between events mean deviation of distance between events overall percentages, standard deviation, and Pairwise Variability Index Pronunciation Acoustic model likelihood scores generated during forced alignment with a native speaker acoustic model the average word-level confidence score of ASR the average difference between the vowel durations in the utterance and vowel-specific means based on a corpus of native speech Grammar Similarity scores of the grammar of the response in ASR with respect to reference response. Vocabulary Use Features about how diverse and sophisticated the vocabulary based on the ASR output.	Linear regression model (with feature selection based on LASSO regression) SpeechRater
4	(Kang and Johnson, 2018)	Fluency	**AM and LM:** KALDI speech recognition engine **Feature selection:** Generative algorithm with k-fold cross-validation	**35 suprasegmental features:** Number of tone units/number of runs; Duration of filled pauses/number of filled pauses; Number of syllables/duration of utterance; Number of syllables/(duration of utterance – duration of silent pauses); Number of tone units with specified relative pitch-tone choice combination per second (Low-rise rate; Mid-rise rate; Low-level rate; Low-fall rate; Mid-fall rate; High-rise-fall rate; High-fall-rise rate) GA selection method was applied to select the best predictive features (11 suprasegmental features)	The boosting ensemble of decision trees

Table 1. (*Continued*).

	Author(s)	Focus	Components of the ASR	Features extracted for scoring	Scoring model
5	(Wang et al., 2018)	Fluency; Linguistic proficiency	**AM:** a speaker adapted Tandem GMM-HMM system a stacked Hybrid system **LM:** Kneser-Ney trigram LM a general English LM **Feature selection:** bottleneck (BN) DNN	A total of 33 features **Audio features:** Energy – mean/ standard deviation; **Fluency features:** Silence – duration mean/ duration standard deviation; Long silence— duration mean; Words – number/ frequency; Phone – duration mean/ duration median; **Linguistic features:** **Parse tree features:** plural common noun/ singular common noun/ general adverb/ general preposition/ article **PoS tag features: f**requency-inverse document frequency features of PoS tags **Pronunciation features** K-L divergence distance	Gaussian processes
6	(Yoon and Bhat, 2018)	Syntactic complexity	**AM:** HMM recognizer, gender independent triphone acoustic model (Yoon and Bhat, 2012) **LM:** four-gram language models **Feature selection:** PoS-based VSMs (Yoon and Bhat, 2012)	Conventional measure: Mean length of clauses (MLC) Mean length of T-units (MLT) Dependent clause ratio (DCC) T-unit complexity (CTU) PoS-based similarity measures: cosmax: the score level with the highest similarity score given the response. cos4: cosine similarity value of the test response with the representative vector of score level i = 4, the highest score level;	Multiple linear regression
7	(Li et al., 2020)	Task completion	AM and LM not specified **Feature selection:** natural language understanding module word2vector model	cosine similarity	Multilayer feed-forward neural network

Table 1. (*Continued*).

	Author(s)	Focus	Components of the ASR	Features extracted for scoring	Scoring model
8	(Fu et al., 2020)	Pronunciation	**Non-native ASR based on ERJ** **AM:** DNN-based acoustic model **LM:** CMU pronouncing dictionary bigram and trigram language models open trigram language model **Feature selection:** GMM-HMM system **Native ASR based on TIMIT** **AM:** DNN-based acoustic model **LM:** CMU pronouncing dictionary trigram language model open trigram language model **Feature selection:** GMM-HMM system	FBANK or MFCC features HMM-based phone log-likelihood score; Word error rate; Reference-free error rate = Comparing the result from the non-native ASR and the native ASR	Linear regression model
9	(Cheng and Wang, 2022)	Fluency; Grammatical proficiency; Task completion	**AM+LM:** Microsoft's local speech recognizer **Feature selection:** Latent Meaning Analysis (LSA) GloVe	**Phonetic features:** Speed of speech Number of voice pauses Number of pronunciation pauses Posterior probability score of pronunciation **Text features:** Total number of words in text Number of nonrepeating words in the text Sum of all syntactic tree depths in text Semantic similarity between text and theme (with LSA) Correct rate of text grammar	BP model CNN + LSTM model
10	(Hayashi et al., 2024)	Fluency	**ASR:** IBM Watson Speech-to-Text **Feature selection:** NLP and speech processing algorithms	token (the length of the utterance) complexity; the number of hesitation/filled pauses; confidence scores.	Random forests

Table 1. (*Continued*).

	Author(s)	Focus	Components of the ASR	Features extracted for scoring	Scoring model
11	(Kang et al., 2024)	Fluency	**ASR:** end-to-end ASR system using a transformer-based encoder–decoder framework **Feature selection:** transcription using the ASR system forced-alignment algorithm determines time-aligned sequences of words and phonemes	**Acoustic features:** Segmental features Intonation Rate	Linear regression Neural network

4. Results

4.1. The construction of the automatic scoring systems

In the reviewed studies, the automatic scoring systems were constructed with ASR models and the scoring models. In this section, the technology or algorithms applied in these two components will be elaborated.

4.1.1. Acoustic modeling in ASR models

ASR models, which directly influence the performance of scoring systems, are an indispensable component in automatic scoring systems. The reviewed studies either utilized pre-existing ASR models or ASR models specifically trained for research purposes.

Regarding the existing ASR models applied in the reviewed studies, Cheng and Wang (2022) employed Microsoft's local speech recognizer, Kang and Johnson (2018) used the KALDI speech recognition engine, and Hayashi et al. (2024) utilized IBM Watson Speech-to-Text model.

Most of the reviewed studies developed their own ASR models for research purposes, typically consisting of separately trained acoustic and language models. Bhat and Yoon (2015), Wang et al. (2018), as well as Yoon and Bhat (2018) all employed the traditional GMM-HMM acoustic model. However, to cater to different research focuses, Bhat and Yoon (2015) used bigram, trigram and four-gram language models; Wang et al. (2018) incorporated a Kenser-Ney trigram language model and a general language model; and Yoon and Bhat (2018) utilized a four-gram language model.

Additionally, the recognition capability of DNN-HMM acoustic models were researched and benchmarked with GMM-HMM acoustic models in studies by Qian et al. (2016), Tao et al. (2016) and Fu et al. (2020). These studies suggested that a higher accuracy in the recognition, or lower Word Error Rate (WER), can be achieved with DNN-HMM acoustic models, ultimately translating into improved performance in scoring systems. Furthermore, Fu et al. (2020) also studied the different settings of the DNN-HMM acoustic model, namely the NNET1 method and three NNET2 methods. Unlike traditional ASR models that require separate training for acoustic and language models separately, more advanced end-to-end ASR models can be trained as a single model. In Kang et al. (2024)'s study, the end-to-end ASR with transformer-based encoder–decoder framework was trained with one set of data only.

4.1.2. The scoring models in automatic scoring systems

A proportion of the scoring models in the reviewed studies were built using linear regression, based on the assumption of a linear relationship between proficiency and features extracted from acoustic signals. Specifically, Fu et al. (2020) and Kang et al. (2024) applied standard linear regression to analyze log-likelihood scores generated by HMM Acoustic Models. Tao et al. (2016) used a similar approach but incorporated LASSO regression for feature selection, while Bhat and Yoon (2015) and Yoon and Bhat (2018) employed multiple linear regression to address their research objectives.

Machine learning algorithms were also incorporated into scoring system modeling, specifically Gaussian Processes, ensemble learning methods and neutral

networks. Gaussian Processes was applied by Wang et al. (2018). For ensemble learning algorithms, Kang and Johnson (2018) utilized decision trees method, while Hayashi et al. (2024) and Qian et al. (2016) both applied the random forest method. Neutral networks were utilized by Li et al. (2020), who incorporated a multilayer feed-forward neural network, and Cheng and Wang (2022), who employed a CNN and an LSTM model, along a BP model.

Regarding the prediction accuracy of the scoring models, all the reviewed studies applied Pearson correlation against the human raters.

4.2. Feature extraction in automatic scoring systems

The features selected for the evaluation are categorized into three types in the present study, namely, phonological competence, linguistic proficiency and task completion. These three categories reflect fluency, complexity, and accuracy as measures of oral proficiency in English as a second language (Housen and Kuiken, 2009), in which fluency measures whether speakers can speak with similar pace like native (Lennon, 1990), complexity reflects the variety of language patterns and accuracy the ability to utter the language without mistakes (Ellis, 2009).

4.2.1. Feature selections for the evaluation of phonological competence

In the studies reviewed, phonological competence was evaluated through assessments of fluency, pronunciation accuracy, and intonation, stress and rhythm.

a) Fluency

Being a major criterion in language assessment, the definition of fluency surprisingly lacks consensus. There is debate on whether it should be considered in a broad sense as synonymous with oral proficiency, and even as overall mastery of a language, or in a narrow sense as just one component of oral proficiency – the uninterpreted flow of speech (Chambers, 1997). Applying the narrower definition, Lennon (1990) quantified fluency using 12 temporal variables related to pauses, run-time, self-correction, and repetition, laying the foundation for most of the automatic evaluation system research that studies fluency. In these systems, fluency is typically conceptualized through a set of acoustic features, namely the speed of the speech, the number of words and its frequency in an uninterrupted chunk, the mean length and frequency of pauses and filled pauses (Bhat and Yoon, 2015; Cheng and Wang, 2022; Hayashi et al., 2024; Kang et al., 2024; Tao et al., 2016; Wang et al., 2018).

b) Pronunciation

While the assessment of fluency, or the uninterrupted flow of speech, can rely solely on the analysis of acoustic vectors extracted from the audio, evaluating pronunciation accuracy requires a benchmark—a model trained on speech corpora, against which the spoken words in the audio can be compared. Wang et al. (2018) compared the differences between the audio inputs from test takers and those of proficient speakers in the model using Kullback-Leibler (K-L) divergece, whereas Tao et al. (2016) and Fu et al. (2020) employed likelihood scores calculated during the forced alignment with the acoustic model trained on corpora of native speech.

c) Intonation, stress and rhythm

Intonation, stress and rhythm have been empirically tested to be distinguishable between proficiency and novice language learners (Anderson-Hsieh et al., 1992). In

automatic assessment systems, these prosody features are quantified using suprasegmental features from the acoustic mode (Kang et al., 2024; Kang and Johnson, 2018; Tao et al., 2016). Specifically, Tao et al. (2016) employed Pairwise Variability Index, the widely used metric for language rhythm quantification, while Kang et al. (2024) incorporated the number of tone units/number of interrupted chunks the number of syllables/duration of utterance, the number of tone units with specified relative pitch-tone choice combination per second (Low-rise rate; Mid-rise rate; Low-level rate; Low-fall rate; Mid-fall rate; High-rise-fall rate; High-fall-rise rate).

4.2.2. Feature selections for the evaluation of linguistic proficiency

In the studies reviewed in this study, the evaluation of linguistic proficiency was achieved by quantifying lexical diversity, grammatical proficiency, and syntactic complexity.

a) Lexical diversity

The measurement of lexical diversity, or vocabulary range, was straightforward. Studies approached it by counting the number of non-repeated words in the ASR output (Cheng and Wang, 2022; Tao et al., 2016).

b) Grammatical proficiency

Applying a similarity measure, Tao et al. (2016) evaluated grammatical proficiency by calculating the similarity score between the grammar in students' utterances and the reference answers. Likewise, Wang et al. (2018) benchmarked candidates' usage of PoS tags against proficient speakers. By applying TF-IDF to 151 PoS tags, the importance of each PoS tag was calculated, and the linguistic features that correlated most with proficient performance were identified: plural common noun, singular common noun, general adverb, general preposition and article (e.g., the, no) (Wang et al., 2018). On the other hand, Cheng and Wang (2022) approached through EASE, an open source composition scoring system, in which 3-gram and 4-gram PoS tags were extracted from Sherlock Holmes' novel collection. The underlying assumption is there is no grammatical mistakes in this published series of novels; therefore, all the PoS tag combinations are the correct combinations of part of speech. If a PoS tag combination extracted from a student's oral delivery cannot be found in the EASE, it is deemed grammatically incorrect (Cheng and Wang, 2022).

c) Syntactic complexity

Besides being a crucial feature in the evaluation of grammatical accuracy, PoS tags have also been employed to assess the syntactic complexity of students' oral deliveries. Bhat and Yoon (2015) used PoS tags to benchmark students' oral deliveries against those of students in four different grade categories, assuming students from the same grade category share similar patterns in their usage of parts of speech. Approaching the task through PoS-based vector space modelling and PoS n-gram language modelling, Bhat and Yoon (2015) found PoS-based vector space modelling to have a better correlation between human and machine grading of students' oral performance. This method was also found to be more effective than traditional measures such as dividing the number of words by the number of clauses, dividing the number of words by the number of T-units, calculating the percentage of dependent clauses relative to the total number of clauses, and counting the number of clauses per T-unit (Yoon and Bhat, 2018). Approaching it differently were Cheng and Wang

(2022), who incorporated syntax trees to evaluate candidates' syntactic complexity in the oral deliveries—the deeper the tree, the more syntactically complex their deliveries.

4.2.3. Feature selections for the evaluation of task completion

The assessment of task completion requires grading systems to evaluate the meaning of utterances, or if the participants have provided off-topic answers to the question. To achieve this, natural language understanding methods, particularly semantic similarity measures, are applied.

In the reviewed studies on evaluating task completion in candidates' oral deliveries, similarity measures were widely adopted. These studies often employed TF-IDF for text vectorization, and Word2Vec, GloVe, LSA or CVA for transforming the ASR output into a lower-dimensional space before applying similarity measurement techniques that benchmarked the utterances against a language model trained on specific topics or score levels (Cheng and Wang, 2022; Li et al., 2020; Qian et al., 2016).

To assess the semantic similarity between candidates' utterance and the given topic of the speaking task, Li et al. (2020) applied Word2vector and cosine similarity. Taking a different approach, Cheng and Wang (2022) vectorized the ASR output using TF-IDF and carried out similarity measurement by training an LSA topic model.

5. Potential research areas and conclusion

The present study reviewed existing empirical studies on the automatic scoring systems of spontaneous L2 English conducted over the past decade, from 2014 to 2024, mapping the components involved in the construction of automatic speech recognition systems and scoring models, as well as the features studied for different evaluation aspects. Overall, the architecture of the scoring system transformed from statistical modelling to deep learning modelling, and in recent years, scholars are exploring the application of transformer modelling. This progression has streamlined model training, as well as improved the accuracy of the grading.

5.1. The development of an automatic scoring model with a holistic approach

In these studies, audio features have been extracted to evaluate students' phonological competence, linguistic proficiency, and task completion, aligning with fluency, complexity, and accuracy as measures of oral proficiency in English as a second language (Housen and Kuiken, 2009). While the studies have successfully developed automatic scoring systems with grades that correlate with human ratings, these systems have typically assessed only one or two evaluation aspects—rather than all three criteria of phonological competence, linguistic proficiency, and task completion. However, each criterion alone is insufficient for building a reliable and effective grading system. For example, if a student delivers a fluent, grammatically accurate response that is off-topic, it cannot be considered a correct answer to the question. Therefore, research is still needed to develop a scoring system that comprehensively evaluates all three criteria.

Another issue in current automatic speaking assessment research is the predominant focus on evaluating spontaneous monologues. Given the crucial communicative aspect of oral language skills, a scoring system that can engage in meaningful conversations with students would more accurately assess their real-world language proficiency. Developing an evaluation system that can handle both spontaneous and interactive dialogue would advance the field, allowing for a more comprehensive assessment of students' ability to respond dynamically in various contexts.

With advancements in large language models (LLMs), speech-based conversational AI models have emerged. These models can not only "listen," "comprehend," and "respond" to unscripted audio inputs but also "listen" while "speaking" simultaneously (Défossez et al., 2024; Ma et al., 2024), facilitating the simulation of natural conversations. This development lays the groundwork for communicative automatic speech evaluation systems.

By incorporating evaluation components into these conversational models—where system responses to students can be pre-configured using prompts—such systems can engage in interactive dialogues, aligning with the communicative essence of language. Beyond assessing traditional criteria—phonological competence, linguistic proficiency, and task completion—these systems could evaluate students' ability to manage real-time conversations, offering a more comprehensive and dynamic approach to assessing language as a tool for communication.

5.2. Study of the effect of the application of these automatic scoring systems

Automatic writing assessment systems have been shown to motivate students (Nunes et al., 2022). With real-time feedback provided by these systems, students who received input from both teachers and the system demonstrated greater persistence in L2 English writing (Franzke et al., 2005; Wilson and Czik, 2016) compared to the control group, who received feedback solely from their teachers. Students expressed that the immediacy of feedback, rather than waiting for comments from teachers, motivated them to write more frequently (Grimes and Warschauer, 2010).

However, findings on whether automatic writing assessment systems improve students' writing skills are mixed. In a comparison study by Wang et al. (2013), students who received feedback from both teachers and the system outperformed the control group, who received feedback only from teachers. Conversely, other studies found no significant difference between the test and control groups (Franzke et al., 2005; Mørch et al., 2017).

Research on the effectiveness of automatic speaking assessment remains limited. Further investigation is needed to determine whether these systems can motivate students in ways similar to automatic writing evaluation systems and to what extent they can enhance users' oral skills. Addressing these questions could help fill this gap and provide deeper insights into the potential of automatic speaking assessment to support language learning and development.

5.3. Conclusion

Given the preliminary nature of automatic scoring systems for spontaneous L2 English and the numerous areas for further exploration, future studies should adopt a more holistic approach to the automatic evaluation of utterances. This approach should encompass not only phonological competence, linguistic proficiency, and task completion but also elements of communicative effectiveness across varied conversational contexts. Additionally, research should investigate the effectiveness of these automatic scoring systems, assessing their impact on students' second language development.

Furthermore, exploring automatic assessment models can contribute to the sustainability of education by creating scalable and accessible tools that reduce the need for extensive human resources. Such systems can also democratize access to language training, providing students worldwide with an accessible method to develop and assess their skills consistently. By supporting a more inclusive and resource-efficient educational environment, these advancements can help create long-term, equitable access to quality language education.

Conflict of interest: The author declares that there are no conflicts of interests.

References

Aksënova, A., Chen, Z., Chiu, C.-C., et al. (2022). Accented Speech Recognition: Benchmarking, Pre-training, and Diverse Data (arXiv:2205.08014). arXiv. https://doi.org/10.48550/arXiv.2205.08014

Alharbi, S., Alrazgan, M., Alrashed, A., et al. (2021). Automatic Speech Recognition: Systematic Literature Review. IEEE Access, 9, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Anderson-Hsieh, J., Johnson, R., Koehler, K. (1992). The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentais, Prosody, and Syllable Structure. Language Learning, 42(4), 529–555. https://doi.org/10.1111/j.1467-1770.1992.tb01043.x

Arslan, L. M., Hansen, J. H. L. (1996). Language accent classification in American English. Speech Communication, 18(4), 353–367. https://doi.org/10.1016/0167-6393(96)00024-6

Benkerzaz, S., Elmir, Y., Dennai, A. (2019). A Study on Automatic Speech Recognition. Journal of Information Technology Review, 10(3). https://doi.org/10.6025/jitr/2019/10/3/77-85

Bhat, S., Yoon, S.-Y. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. Speech Communication, 67, 42–57. https://doi.org/10.1016/j.specom.2014.09.005

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18(4), 467–480. Retrieved from https://aclanthology.org/J92-4003/

Chambers, F. (1997). What do we mean by fluency? System, 25(4), 535–544. https://doi.org/10.1016/S0346-251X(97)00046-8

Chen, M., Zechner, K. (2011). Computing and Evaluating Syntactic Complexity Features for Automated Scoring of Spontaneous Non-Native Speech. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; June 2011; Portland, OR, USA. pp. 722–731

Chen, Y., Hu, J., Zhang, X. (2019). Sell-corpus: An Open Source Multiple Accented Chinese-english Speech Corpus for L2 English Learning Assessment. In: Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 12–17 May 2019; Brighton, UK. 7425–7429. https://doi.org/10.1109/ICASSP.2019.8682612

Cheng, J., Chen, X., Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. Speech Communication, 73, 14–27. https://doi.org/10.1016/j.specom.2015.07.006

Cheng, Z., Wang, Z. (2022). Automatic Scoring of Spoken Language Based on Basic Deep Learning. Scientific Programming, 2022, 1–14. https://doi.org/10.1155/2022/6884637

Cucchiarini, C., Strik, H., Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. The Journal of the Acoustical Society of America, 107(2), 989–999. https://doi.org/10.1121/1.428279

Cucchiarini, C., Strik, H., Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. The Journal of the Acoustical Society of America, 111(6), 2862–2873. https://doi.org/10.1121/1.1471894

Défossez, A., Mazaré, L., Orsini, M., et al. (2024). Moshi: A speech-text foundation model for real-time dialogue. arXiv. https://doi.org/10.48550/arXiv.2410.00037

Deng, Y., Li, X., Kwan, C., Raj, B., & Stern, R. (2007). Continuous feature adaptation for non-native speech recognition. International Journal of Computer and Information Engineering, 1(6), 1701–1707. https://doi.org/10.5281/zenodo.1329829

Dufour, R., Estève, Y., Deléglise, P. (2014). Characterizing and detecting spontaneous speech: Application to speaker role recognition. Speech Communication, 56, 1–18. https://doi.org/10.1016/j.specom.2013.07.007

Ellis, R. (2009). Task-based language learning and teaching (7th print). Oxford University Press.

Fendji, J. L. K. E., Tala, D. C. M., Yenke, B. O., Atemkeng, M. (2022). Automatic Speech Recognition Using Limited Vocabulary: A Survey. Applied Artificial Intelligence, 36(1), 2095039. https://doi.org/10.1080/08839514.2022.2095039

Franzke, M., Kintsch, E., Caccamise, D., et al. (2005). Summary Street®: Computer Support for Comprehension and Writing. Journal of Educational Computing Research, 33(1), 53–80. https://doi.org/10.2190/DH8F-QJWM-J457-FQVB

Fu, J., Chiba, Y., Nose, T., Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. Speech Communication, 116, 86–97. https://doi.org/10.1016/j.specom.2019.12.002

Gabler, P., Geiger, B. C., Schuppler, B., Kern, R. (2023). Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition. Information, 14(2), 137. https://doi.org/10.3390/info14020137

Ge, Z. (2015). Improved accent classification combining phonetic vowels with acoustic features. In: Proceedings of the 2015 8th International Congress on Image and Signal Processing (CISP); 14–16 October 2015; Shenyang, China. pp. 1204–1209. https://doi.org/10.1109/CISP.2015.7408064

Georgakis, C., Petridis, S., Pantic, M. (2016). Discrimination Between Native and Non-Native Speech Using Visual Features Only. IEEE Transactions on Cybernetics, 46(12), 2758–2771. https://doi.org/10.1109/TCYB.2015.2488592

Gerosa, M., Giuliani, D., & Narayanan, S. (2006). Acoustic analysis and automatic recognition of spontaneous children's speech. Interspeech 2006, 519–522. https://doi.org/10.21437/Interspeech.2006-519

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. Journal of Technology, Learning, and Assessment, 8(6). Retrieved June 20, 2024, from http://www.jtla.org

Hayashi, Y., Kondo, Y., Ishii, Y. (2024). Automated speech scoring of dialogue response by Japanese learners of English as a foreign language. Innovation in Language Learning and Teaching, 18(1), 32–46. https://doi.org/10.1080/17501229.2023.2217181

Heeman, P. A., & Allen, J. F. (1999). Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. Computational Linguistics, 25(4), 527–572. Retrieved from https://aclanthology.org/J99-4003/

Housen, A., Kuiken, F. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. Applied Linguistics, 30(4), 461–473. https://doi.org/10.1093/applin/amp048

Johnson, M., Charniak, E. (2004). A TAG-based noisy channel model of speech repairs. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics; 21–26 July 2004; Stroudsburg PA USA. p. 33-es. https://doi.org/10.3115/1218955.1218960

Kang, B. O., Jeon, H., Lee, Y. K. (2024). AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation. ETRI Journal, 46(1), 48–58. https://doi.org/10.4218/etrij.2023-0322

Kang, O., Johnson, D. (2018). The roles of suprasegmental features in predicting English oral proficiency with an automated system. Language Assessment Quarterly, 15(2), 150–168. https://doi.org/10.1080/15434303.2018.1451531

Kat L. W., Fung, P. (1999). Fast accent identification and accented speech recognition. In: Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258); 15–19 March 1999; Phoenix, AZ, USA. pp. 221–224. https://doi.org/10.1109/ICASSP.1999.758102

Knill, K. M., Gales, M. J. F., Manakul, P. P., Caines, A. P. (2019). Automatic Grammatical Error Detection of Non-native Spoken Learner English. In: Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 12–17 May 2019; Brighton, UK. pp. 8127–8131. https://doi.org/10.1109/ICASSP.2019.8683080

Kobayashi, Y., & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. Journal of Pan-Pacific Association of Applied Linguistics, 20(1), 55–73. Retrieved from https://eric.ed.gov/?id=EJ1110804

Lease, M., Johnson, M., Charniak, E. (2006). Recognizing disfluencies in conversational speech. IEEE Transactions on Audio, Speech and Language Processing, 14(5), 1566–1573. https://doi.org/10.1109/TASL.2006.878269

Lee, H.-S., Chen, P.-Y., Cheng, Y.-F., et al. (2022). Speech-enhanced and Noise-aware Networks for Robust Speech Recognition. arXiv. http://arxiv.org/abs/2203.13696

Lennon, P. (1990). Investigating Fluency in EFL: A Quantitative Approach. Language Learning, 40(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Li, K.-C., Chang, M., Wu, K.-H. (2020). Developing a Task-Based Dialogue System for English Language Learning. Education Sciences, 10(11), 306. https://doi.org/10.3390/educsci10110306

Li, S., Ouyang, B., Liao, D., et al. (2021). End-To-End Multi-Accent Speech Recognition with Unsupervised Accent Modelling. In: Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 6–11 June 2021; Toronto, ON, Canada. pp. 6418–6422. https://doi.org/10.1109/ICASSP39728.2021.9414833

Liu, J., Wumaier, A., Fan, C., Guo, S. (2023). Automatic Fluency Assessment Method for Spontaneous Speech without Reference Text. Electronics, 12(8), 1775. https://doi.org/10.3390/electronics12081775

Livescu, K., Glass, J. (2000). Lexical modeling of non-native speech for automatic speech recognition. In: Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100); 5–9 June 2000; Istanbul, Turkey. 1683–1686. https://doi.org/10.1109/ICASSP.2000.862074

Ma, Z., Song, Y., Du, C., et al. (2024). Language Model Can Listen While Speaking. arXiv. http://arxiv.org/abs/2408.02622

Miao, Y., Metze, F. (2017). End-to-End Architectures for Speech Recognition. In: Watanabe, S., Delcroix, M., Metze, F., Hershey, J. R. (editors). New Era for Robust Speech Recognition, Springer International Publishing. pp. 299–323. https://doi.org/10.1007/978-3-319-64680-0_13

Molenaar, B., Tejedor-Garcia, C., Cucchiarini, C., Strik, H. (2023). Automatic Assessment of Oral Reading Accuracy for Reading Diagnostics. INTERSPEECH, 2023, 5232–5236. https://doi.org/10.21437/Interspeech.2023-1681

Mørch, A., Engeness, I., Cheung, K.-W. (2017). EssayCritic: Writing to learn with a knowledge-based design critiquing system. Educational Technology & Society, 20(2), 213–223.

Nakamura, M., Iwano, K., Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Computer Speech & Language, 22(2), 171–184. https://doi.org/10.1016/j.csl.2007.07.003

Nunes, A., Cordeiro, C., Limpo, T., Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. Journal of Computer Assisted Learning, 38(2), 599–620. https://doi.org/10.1111/jcal.12635

Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, n71. https://doi.org/10.1136/bmj.n71

Qian, Y., Wang, X., Evanini, K., Suendermann-Oeft, D. (2016). Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment. Interspeech, 2016, 3122–3126. https://doi.org/10.21437/Interspeech.2016-291

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? Proceedings of the IEEE, 88(8), 1270–1278. https://doi.org/10.1109/5.880083

Sainath, T. N., Weiss, R. J., Wilson, K. W., et al. (2017). Raw Multichannel Processing Using Deep Neural Networks. In: Watanabe, S., Delcroix, M., Metze, F., Hershey, J. R. (Eds.). New Era for Robust Speech Recognition, Springer International Publishing. pp. 105–133. https://doi.org/10.1007/978-3-319-64680-0_5

Saon, G., Chien, J.-T. (2012). Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances. IEEE Signal Processing Magazine, 29(6), 18–33. https://doi.org/10.1109/MSP.2012.2197156

Shriberg, E. E. (1999). Phonetic consequences of speech disfluency. Proceedings of the International Congress of Phonetic Sciences, 1(2), 619–622. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0619.pdf

Tamazin, M., Gouda, A., Khedr, M. (2019). Enhanced Automatic Speech Recognition System Based on Enhancing Power-Normalized Cepstral Coefficients. Applied Sciences, 9(10), 2166. https://doi.org/10.3390/app9102166

Tao, J., Ghaffarzadegan, S., Chen, L., Zechner, K. (2016). Exploring deep learning architectures for automatically grading non-native spontaneous speech. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 20–25 March 2016; Shanghai, China. pp. 6140–6144. https://doi.org/10.1109/ICASSP.2016.7472857

Trentin, E., Gori, M. (2001). A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing, 37(1–4), 91–126. https://doi.org/10.1016/S0925-2312(00)00308-8

Van Bergem, D. R. (1995). Perceptual and acoustic aspects of lexical vowel reduction, a sound change in progress. Speech Communication, 16(4), 329–358. https://doi.org/10.1016/0167-6393(95)00003-7

Wang, Y., Gales, M. J. F., Knill, K. M., et al. (2018). Towards automatic assessment of spontaneous spoken English. Speech Communication, 104, 47–56. https://doi.org/10.1016/j.specom.2018.09.002

Wang, Y.-J., Shang, H.-F., Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. Computer Assisted Language Learning, 26(3), 234–257. https://doi.org/10.1080/09588221.2012.655300

Wills, S., Bai, Y., Tejedor-Garcia, C., et al. (2023). Automatic Speech Recognition of Non-Native Child Speech for Language Learning Applications. Slate. https://doi.org/10.4230/OASICS.SLATE.2023.11

Wilson, J., Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. Computers & Education, 100, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Yoon, S. Y., & Bhat, S. (2012). Assessment of ESL learners' syntactic competence based on similarity measures. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 600–608). Association for Computational Linguistics. https://doi.org/10.3115/v1/D12-1055

Yoon, S.-Y., Bhat, S. (2018). A comparison of grammatical proficiency measures in the automated assessment of spontaneous speech. Speech Communication, 99, 221–230. https://doi.org/10.1016/j.specom.2018.04.003

Yu, D., Deng, L. (2015). Automatic Speech Recognition: A Deep Learning Approach. Springer London. https://doi.org/10.1007/978-1-4471-5779-3

Zechner, K., Evanini, K., Yoon, S.-Y., et al. (2014). Automated scoring of speaking items in an assessment for teachers of English as a Foreign Language. In: Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications; June 2014; Baltimore, Maryland. pp. 134–142. https://doi.org/10.3115/v1/W14-1816