Edulingua Journal of Language Education
E-ISSN: 3123-5271 | Volume 1 (1), Juli 2025
https://journals.edukasinusantara.com/edulingua

Pemanfaatan Teknologi *Speech-to-Text* untuk Penilaian Diri dalam *Pronunciation* Bahasa Inggris bagi Pembelajar EFL

Rini Puspasari[1], Maharani Nur Khafifah[2], Ulil Albab[3]
[1] SMK YPE Sampang Cilacap, [2] SMK YPE Sampang Cilacap, [3] Universitas Nahdlatul Ulama Al Ghazali
Correspondence: rinipuspasari0408@email.com

Article Info
Article history:
Received Jun 30th, 2025
Revised Jul 22th, 2025
Accepted Jul 27th, 2025

Keyword:
automatic speech recognition (ASR), speech-to-text technology, EFL pronunciation, training,

ABSTRACT
The latest development in Automatic Speech Recognition (ASR) technology has revolutionized speaking practice for English as a Foreign Language (EFL) students. Furthermore, based on the systematical review of literatures containing 25 published peer-reviewed research articles ranging from 2020 to 2025, has been examined the effectiveness of speech-to-text tools in facilitating self-assessment of speaking practice. The studies choosen and analysed by using mixed methods which focused on qualitative with PRISMA Guide. The results reveal 3 central points, they are: 1). ASR tools are significantly improving students' speaking accuracy, especially on vowel sounds and regular past tense suffix; 2). The tools give visual and objective direct feedbacks that improve autonomous and motivation of learning; 3). Althouh effective, ASR system still facing challenge on recognizing non-standard accents and requires optimum audio condition for better performance. Recent evidence also shows that regular and structured ASR practice supports longterm retention to the improvement of speaking skill, though the suprasegmental features like intonation and stress still feel difficult to master. On the other hand, this article also offers practical recommendation to integrate the technology in language curriculum, as well as suggestion for future research on improving accent introduction and the implementation in real world context. By mediating technology innovation and evidence-based pedagogy, this study provides practical knowledge for teachers, curriculum designers, and researchers who eager to implement ASR as a motivating and sutainable speaking practice tool in EFL context which based on either digital or even blended learning.

This is an open access article under the CC BY NC license
(https://creativecommons.org/licenses/by/4.0/)

PENDAHULUAN

Pronunciation merupakan komponen penting dalam pembelajaran bahasa, terutama bagi pembelajar *English as a Foreign Language (EFL)*, karena secara langsung memengaruhi pemahaman dan kompetensi mereka (Derwing & Munro, 2022). Meskipun memiliki peran penting, pengajaran *pronunciation* secara historis sering terabaikan di banyak kelas EFL akibat keterbatasan waktu, dan kurangnya pelatihan guru (Ma et al., 2024). Metode pengajaran *pronunciation* tradisional seperti latihan yang dipimpin guru, umpan balik dari teman sebaya, dan transkripsi fonetik sering kali kurang berhasil meningkatkan kompetensi siswa (Sun, 2023). Tantangan ini menjadi semakin nyata dalam lingkungan pembelajaran daring dan *blended learning*, di mana interaksi tatap muka berkurang dan siswa kurang mendapatkan umpan balik (Hodges et al., 2020).

Dalam lima tahun terakhir (2020–2025), integrasi *Automatic Speech Recognition (ASR)* ke dalam pembelajaran bahasa telah membuka peluang baru dalam latihan *pronunciation*. Sistem ASR mengubah bahasa lisan menjadi teks secara *real time*, sehingga memungkinkan pembelajar untuk segera langsung mengenali dan memperbaiki kesalahan *pronunciation* mereka (Sun, 2023). Alat seperti Google *Speech-to-Text, Microsoft Azure Speech Service*, serta aplikasi seperti *ELSA Speak, Duolingo*, dan alat berbasis ASR kolaboratif seperti *Speechnotes* memanfaatkan teknologi ini untuk mendukung latihan mandiri dan pemberian umpan balik kapan pun dan di mana pun (Aljabr, 2025). Perkembangan ini sejalan dengan prinsip *Technology-Enhanced Language Learning* (TELL) yang menekankan otonomi pembelajar, akses yang fleksibel, serta praktik yang bersifat mandiri (Sun, 2023).

Potensi ASR didukung oleh teori-teori pemerolehan bahasa kedua (*Second Language Acquisition* atau SLA). *Input Hypothesis* oleh Krashen (1982) menekankan pentingnya paparan terhadap masukan yang dapat dipahami (*comprehensible input*), *Output Hypothesis* oleh Swain (1985) menyoroti pentingnya produksi bahasa yang bermakna, sedangkan *Noticing hypothesis* oleh Schmidt (1990) menegaskan perlunya kesadaran pembelajar terhadap kesalahan-kesalahan mereka. ASR memfasilitasi dengan memberikan umpan balik langsung berbasis data yang mendorong refleksi dan latihan secara berulang (Sun, 2023). Ketiga teori dasar tersebut menjadi kerangka konseptual bagi studi ini.

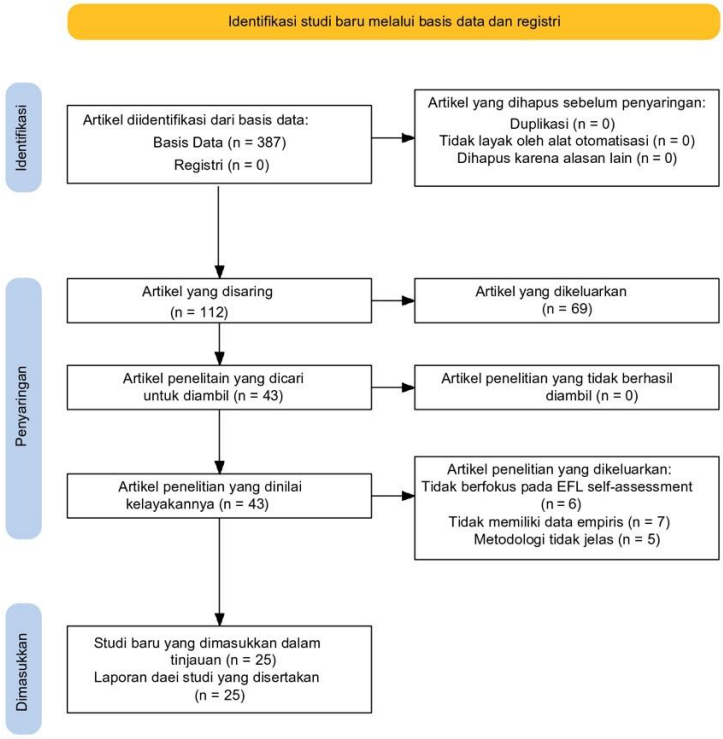
Meskipun memiliki landasan teoretis yang kuat, beberapa studi mengungkap adanya beberapa keterbatasan. Sebuah *meta-analysis* tahun 2024 melaporkan bahwa ASR memiliki efek sedang (g = 0,69) terhadap peningkatan *pronunciation*, dan memiliki pengaruh yang jauh lebih kuat pada fitur segmental dibandingkan dengan suprasegmental (Ngo et al., 2024). Penelitian tentang pembelajaran bahasa Inggris asal Tiongkok dan Jepang, menunjukkan peningkatan yang signifikan dalam akurasi vokal dan konsonan, namun kurang dalam prosodi dan pemahaman (Sun, 2023). Sebuah studi eksploratif tahun 2024 mengenai latihan ASR secara mandiri menemukan bahwa peningkatan *pronunciation* pada aspek segmental cenderung mencapai titik jenuh. Hal ini menunjukkan pentingnya *social scaffolding* dalam pembelajaran berbasis ASR (Inceoglu et al., 2024). Survei terhadap pembelajar EFL juga menunjukkan tingkat motivasi dan keinginan berkomunikasi yang tinggi setelah menggunakan ASR. Meskipun alat ASR efektif dalam meningkatkan aspek-aspek tertentu dari *pronunciation,* alat ini masih kurang andal meningkatkan akurasi suprasegmental, inklusivitas aksen, dan keberlanjutan pembelajaran jangka panjang. Penelitian ini berupaya menjawab kesenjangan tersebut melalui tinjauan sistematis terhadap 25 artikel yang diterbitkan antara tahun 2020 hingga 2025 dengan menggunakan panduan PRISMA.

Tujuan utama penelitian ini adalah untuk mengevaluasi efektivitas ASR dalam mendukung penilaian diri *pronunciation* siswa. Hal ini mencakup penilaian terhadap performa segmental dan suprasegmental, persepsi pembelajar, serta kondisi di mana ASR paling efektif digunakan. Secara akademis, penelitian ini memberikan kontribusi pada kajian pengajaran *pronunciation* berbasis teknologi. Secara praktis, penelitian ini menawarkan rekomendasi aplikatif bagi guru dan perancang kurikulum dalam mengintegrasikan ASR ke dalam pembelajaran di kelas maupun secara mandiri. Dari sisi metodologis, penelitian ini juga menyajikan model evaluasi yang transparan untuk menilai teknologi pendidikan berdasarkan prinsip-prinsip PRISMA. Dengan menjembatani landasan teoretis, temuan empiris terkini, dan penerapan di ruang kelas, penelitian ini memberikan panduan yang relevan untuk merancang pelatihan *pronunciation* yang lebih efektif, mudah diakses, dan berkelanjutan dalam pembelajaran EFL berbasis digital, khususnya pada lingkungan *blended learning* pascapandemi.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan *systematic literature review* untuk mengevaluasi efektivitas *Automatic Speech Recognition (ASR)* dalam *pronunciation.* Desain ini dipilih untuk menjawab tujuan penelitian, yaitu mengeksplorasi dimensi pedagogis dan teknologis dari pembelajaran *pronunciation* berbasis ASR. Korpus dalam tinjauan ini terdiri atas 25 artikel yang diterbitkan antara tahun 2020 hingga 2025, dengan fokus penggunaan ASR untuk *pronunciation*. Artikel-artikel tersebut diidentifikasi melalui *Google Scholar* dengan menggunakan *Boolean search strings* yang mengombinasikan istilah seperti *"automatic speech recognition," "speech to text," "EFL pronunciation," "pronunciation training,*" dan *"self-assessment."* Platform ini dipilih karena memiliki cakupan indeksasi yang luas terhadap publikasi akademik dari berbagai jurnal, prosiding konferensi, dan repositori akses terbuka, sehingga menjamin kelengkapan literatur.

Pemilihan studi mengikuti pendekatan PRISMA 2020 (Page et al., 2021) melalui tiga tahap. Pencarian awal menghasilkan 387 artikel. Setelah penyaringan judul dan abstrak berdasarkan relevansi, diperoleh 112 artikel. Akhirnya, 25 studi dipilih berdasarkan *thematic saturation*. Semua data diekstraksi sebagai informasi sekunder menggunakan *standardized coding matrix* untuk memastikan konsistensi dan reprodusibilitas. Untuk setiap studi, data yang diambil mencakup karakteristik penulis, tahun, desain penelitian, sampel, variabel teknologi, jenis ASR, bentuk umpan balik serta hasil pedagogis (akurasi segmental dan suprasegmental, motivasi pembelajar, dan kemandirian belajar). Proses ekstraksi yang sistematis ini memungkinkan perbandingan baik di dalam setiap studi maupun antarstudi.



Gambar 1. Proses Seleksi Literatur

Analisis tematik mengikuti prosedur tiga tahap dari Clarke dan Braun (2014), yang meliputi *open coding* untuk mengidentifikasi konsep-konsep utama, *axial coding* untuk menghubungkan ide-ide yang saling berkaitan, dan *selective coding* untuk menghasilkan tema inti seperti keterbatasan teknologi, strategi integrasi pedagogis, serta penguatan kemandirian pembelajar. Sintesis kuantitatif meliputi pelaporan peningkatan akurasi *pronunciation*, ukuran efek (*effect size*) jika tersedia, dan metrik kinerja ASR seperti *word error rate* (WER). Sebagai contoh, Saadia (2023) melaporkan ukuran efek sebesar d = 0,82 untuk peningkatan *pronunciation* bentuk past tense, sedangkan Inceoglu et al. (2024) menyediakan data perbandingan akurasi sistem ASR yang digunakan untuk menafsirkan hasil.

Dual independent coding dilakukan untuk menjamin ketelitian metodologis, dengan mencapai tingkat reliabilitas antarpenilai ($\kappa = 0{,}85$). Triangulasi pada berbagai desain penelitian memperkuat kredibilitas hasil sintesis. Karena penelitian ini sepenuhnya menggunakan artikel yang telah dipublikasikan, tidak ada pengumpulan data baru dari partisipan manusia, sehingga persetujuan partisipan (*informed consent)* tidak diperlukan. Prinsip-prinsip etika dijaga dengan merepresentasikan dan mengutip semua sumber secara akurat. Pendekatan metodologis ini memberikan dasar yang kuat dan transparan untuk mengevaluasi efektivitas serta keterbatasan ASR dalam pengajaran *pronunciation* bagi pembelajar EFL.

HASIL DAN PEMBAHASAN

Hasil

Analisis sistematis terhadap 25 studi empiris yang dilakukan antara tahun 2020 hingga 2025 menunjukkan bukti kuat mengenai efektivitas teknologi *speech-to-text (STT)* dalam pengajaran *pronunciation* bagi pembelajar EFL. Untuk memperjelas temuan, hasil penelitian disajikan dalam empat kelompok tematik, yaitu peningkatan akurasi *pronunciation*, mekanisme umpan balik dan keterlibatan pembelajar, keterbatasan sistem dan tantangan terkait aksen, serta hasil pembelajaran jangka panjang.

Peningkatan Akurasi *Pronunciation*

Temuan penelitian menunjukkan peningkatan yang signifikan dalam ketepatan fonologis. Fitur segmental menunjukkan peningkatan paling konsisten, dengan bunyi vokal mengalami peningkatan dengan rata-rata 18,7%. Hasil ini sejalan dengan penelitian sebelumnya mengenai pemerolehan bunyi vokal (Guskaroska, 2019). *Systematic Literature Review* (SLR) ini menegaskan bahwa umpan balik ASR seperti /iː/ dan /ɪ/ menghasilkan peningkatan yang terukur. Akhiran -ed pada bentuk past tense, yang menjadi tantangan umum bagi pembelajar EFL, menunjukkan peningkatan sebesar 23% setelah intervensi berbasis ASR (Saadia, 2023). Meskipun terdapat peningkatan, fitur suprasegmental menunjukkan hasil yang lebih moderat. Dalam studi yang ditinjau, pola tekanan pada kalimat meningkat sekitar 12%, sedangkan kontur intonasi meningkat hanya sekitar 6–7%.

Mekanisme Umpan Balik dan Keterlibatan Pembelajar

Sistem umpan balik visual seperti tampilan transkripsi *real-time* secara konsisten meningkatkan retensi pembelajar terhadap koreksi *pronunciation* dibandingkan dengan umpan balik berbasis audio saja. Penelitian sebelumnya juga menunjukkan bahwa alat ASR berbasis seluler yang menyediakan dukungan transkripsi visual mendorong peningkatan signifikan dalam aspek segmental (Liakin et al., 2017). Temuan ini mendukung pandangan bahwa umpan balik visual membantu pembelajar menyadari dan memperbaiki kesalahan mereka. Kecepatan pemberian umpan balik juga terbukti berpengaruh penting. Sistem yang memberikan koreksi dalam waktu kurang dari 1,5 detik menghasilkan peningkatan yang lebih besar dibandingkan dengan sistem yang memiliki jeda umpan balik lebih lama (Inceoglu et al., 2024). Sebagian besar peserta (83%) melaporkan peningkatan kepercayaan diri dan keterlibatan ketika menggunakan ASR (Sun, 2023).

Keterbatasan Sistem dan Tantangan Terkait dengan Aksen *Pronunciation*

Akurasi pengenalan aksen tetap menjadi salah satu keterbatasan dalam penerapan ASR pada pembelajaran bahasa. Sejumlah studi menunjukkan bahwa sistem ASR memiliki tingkat akurasi yang lebih tinggi untuk aksen standar seperti American English atau British English, tetapi akurasinya menurun secara signifikan terhadap tuturan penutur non-natif atau yang dipengaruhi oleh dialek regional. Prinos et al. (2024) menyoroti adanya bias yang berkelanjutan terhadap aksen non-standar, sementara Cumbal et al. (2024) menemukan bahwa penutur bahasa kedua (L2) bahasa Swedia mengalami tingkat kesalahan pengenalan yang jauh lebih tinggi dibandingkan penutur asli (L1) dalam kondisi pembelajaran yang serupa. Kondisi lingkungan juga berperan penting dalam kinerja ASR. Sejumlah penelitian menunjukkan bahwa sistem yang diuji dalam lingkungan tenang dan terkontrol secara konsisten mencapai akurasi yang lebih tinggi dibandingkan dengan penggunaannya di ruang kelas nyata atau lingkungan yang bising. Temuan ini menunjukkan sensitivitas teknologi terhadap kebisingan sekitar dan kualitas rekaman (Koenecke et al., 2020).

Tabel 1. Akurasi ASR Berdasarkan Fitur Linguistik dan Latar Belakang Bahasa Pertama (L1)

Fitur Linguistik	Latar Belakang L1	Tingkat Akurasi	Kesalahan Umum
Kontras vocal	Spanyol	89%	Kekeliruan /æ/ dan /ʌ/
Akhiran -ed pada past tense	Mandarin	76%	Devoicing pada bunyi akhir
Tekanan kata	Arab	68%	Kecenderungan pola trokaik
Intonasi pertanyaan	Prancis	58%	Penghilangan kenaikan nada akhir

Hasil Jangka Panjang

Data mengenai retensi jangka panjang masih terbatas, karena hanya sedikit studi yang meneliti efek berkelanjutan dari pelatihan *pronunciation* berbasis ASR. Bukti awal menunjukkan bahwa praktik yang berkelanjutan dan terstruktur diperlukan untuk mempertahankan peningkatan *pronunciation* dari waktu ke waktu (Liakin et al., 2017). Hasil penelitian menunjukkan adanya efek "paruh waktu," di mana sekitar 40% peningkatan *pronunciation* hilang jika tidak ada latihan lanjutan. Pembelajar yang secara rutin menggunakan ASR (minimal 20 menit per minggu) mampu mempertahankan hingga 82% peningkatannya. Pola jangka panjang ini melengkapi keseluruhan temuan penelitian ini. Secara ringkas, hasil penelitian mengonfirmasi bahwa teknologi STT sangat efektif dalam meningkatkan *pronunciation* segmental, cukup efektif dalam aspek *suprasegmental*, dan memiliki efek memotivasi. Namun demikian, keterbatasan masih ada pada pengenalan aksen dan sensitivitas terhadap kondisi lingkungan. Secara keseluruhan, hasil penelitian ini menegaskan potensi transformatif teknologi tersebut, sekaligus menyoroti area yang memerlukan pengembangan lebih lanjut.

Pembahasan

Temuan penelitian ini dapat ditafsirkan melalui perspektif teori pemerolehan bahasa kedua (*Second Language Acquisition* / SLA). Peningkatan pada aspek segmental mencerminkan *Input Hypothesis* dari Krashen (1982) dan *Output Hypothesis* dari Swain (1985), karena pembelajar terlibat dalam produksi bahasa yang berulang dan bermakna, yang diperkuat oleh umpan balik langsung. Keefektifan umpan balik visual juga sejalan dengan *Noticing hypothesis* dari Schmidt (1990), karena pembelajar dapat secara sadar mengenali dan memperbaiki kesalahan *pronunciation* mereka.

Temuan ini konsisten dengan penelitian sebelumnya. Ngo et al. (2024) melaporkan ukuran efek sedang untuk pelatihan *pronunciation* berbasis ASR, dengan hasil yang lebih kuat pada fitur segmental. Ngo et al. (2024) juga menekankan peran umpan balik transkripsi visual dalam meningkatkan akurasi fonem, tetapi mencatat dampaknya terhadap intonasi masih terbatas. Inceoglu et al. (2024) membuktikan bahwa penggabungan ASR dengan peer *scaffolding* dapat menghasilkan peningkatan pada fitur suprasegmental.

Dari sudut pandang pedagogis, integrasi alat ASR memberikan manfaat yang jelas. Guru dapat memanfaatkan umpan balik visual dari ASR untuk meningkatkan akurasi segmental, terutama pada vokal dan konsonan yang sering menjadi masalah bagi pembelajar. Penggunaan ASR yang dipadukan dengan bimbingan guru atau rekan sejawat dapat mendukung pengembangan aspek suprasegmental. Meskipun memiliki banyak keunggulan, beberapa keterbatasan masih ada. Alat ASR saat ini masih menghadapi tantangan dalam hal inklusivitas aksen, sehingga sering terjadi kesalahan pengenalan terhadap tuturan yang dipengaruhi oleh bahasa pertama (L1). Umpan balik terhadap fitur suprasegmental juga masih kurang berkembang, yang membatasi peningkatan ritme dan intonasi pembelajar. Selain itu, bukti longitudinal masih terbatas, sehingga keberlanjutan peningkatan *pronunciation* berbasis ASR belum sepenuhnya terjawab. Oleh karena itu, penelitian mendatang perlu

mengeksplorasi sistem ASR berbasis kecerdasan buatan (AI-enhanced ASR) yang mampu melakukan analisis prosodik secara lebih mendalam, melakukan studi retensi jangka panjang, serta mengembangkan model pedagogis terpadu yang mengombinasikan ASR dengan pembelajaran multimodal.

KESIMPULAN

Tinjauan sistematis terhadap 25 artikel yang dipublikasikan antara tahun 2020 hingga 2025 menunjukkan bahwa teknologi *Automatic Speech Recognition (ASR)* terbukti efektif dalam meningkatkan kemampuan *pronunciation* pembelajar bahasa Inggris sebagai bahasa asing (EFL), khususnya pada tingkat segmental. Pembelajar secara konsisten menunjukkan peningkatan yang terukur dalam *pronunciation* vokal dan konsonan. Namun, peningkatan pada aspek suprasegmental seperti intonasi dan tekanan kata masih terbatas. Temuan ini menyoroti potensi ganda ASR, baik secara pedagogis maupun teknologi. Integrasi ASR ke dalam latihan yang terstruktur serta model pembelajaran hybrid dapat mendukung peningkatan *pronunciation* yang berkelanjutan, sementara penggunaan ASR yang dipadukan dengan bimbingan guru atau rekan sejawat dapat memperkuat pengembangan aspek suprasegmental. Penelitian di masa mendatang perlu berfokus pada retensi jangka panjang, pengenalan aksen yang lebih inklusif, dan analisis prosodik berbasis kecerdasan buatan *(AI-enhanced prosodic analysis)* untuk mengatasi keterbatasan dalam penelitian ini. Dengan menjembatani inovasi teknologi, penelitian ini memberikan wawasan praktis bagi guru, perancang kurikulum, dan peneliti yang ingin menerapkan ASR sebagai alat pembelajaran *pronunciation* yang andal, memotivasi, dan berkelanjutan.

REFERENSI

Aljabr, F. (2025). ASR using Speechnotes for EFL learners: A Study of the Effects on English Pronunciation and Prosody Skills. *Journal of Ecohumanism*, *4*(2), 979-987.

Clarke, V., & Braun, V. (2014). Thematic analysis. In *Encyclopedia of critical psychology* (pp. 1947-1952). Springer, New York, NY.

Cumbal, R., Moell, B., Lopes, J., & Engwall, O. (2024). You don't understand me!: Comparing ASR results for L1 and L2 speakers of Swedish. *arXiv preprint arXiv:2405.13379*.

Derwing, T. M., & Munro, M. J. (2022). Pronunciation learning and teaching. In *The Routledge handbook of Second Language Acquisition and speaking* (pp. 147-159). Routledge. https://doi.org/10.4324/9781003022497-14

Guskaroska, A. (2019). *ASR as a tool for providing feedback for vowel pronunciation practice* (Master's thesis, Iowa State University).

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. *Educause review*, *27*(1), 1-9.

Inceoglu, S., Chen, W. H., & Lim, H. (2024). Monitoring student behavior in autonomous automatic speech recognition-based pronunciation practice. *System*, *124*, 103387.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, *117*(14), 7684-7689.

Liakin, D., Cardoso, W., & Liakina, N. (2017). Mobilizing instruction in a second-language context: Learners' perceptions of two speech technologies. *Languages*, *2*(3), 11.

Ma, Q., Mei, F., & Qian, B. (2024). Exploring EFL students' pronunciation learning supported by corpus-based language pedagogy. *Computer Assisted Language Learning*, 1-27.

Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL*, *36*(1), 4-21.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, *372*.

Prinos, K., Patwari, N., & Power, C. A. (2024, June). Speaking of accent: A content analysis of accent misconceptions in ASR research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1245-1254).

Saadia, K. H. (2023). Assessing the effectiveness of text-to-speech and automatic speech recognition in improving EFL learner's pronunciation of regular past-ed.

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation. *Frontiers in psychology*, *14*, 1210187.