Bit-Tech (Binary Digital - Technology)

Vol.7, No.2, Desember 2024
Available online at: http://jurnal.kdi.or.id/index.php/bt

Teknologi Pengenalan Suara tentang Metode, Bahasa dan Tantangan: *Systematic Literature Review*

Hartana Wijaya[1)] *

[1)]*Fakutas Sains dan Teknologi, Universitas Buddhi Dharma*
Jl. Imam Bonjol No. 41 Karawaci, Tangerang, Indonesia

[1)]hartana.wijaya@ubd.ac.id

Article history:

Received 05 Des 2024;
Revised 09 Des 2024;
Accepted 19 Des 2024;
Available online 27 Des 2024

Keywords:

Deep learning
Machine learning
Pengenalan suara
PRISMA
SLR

Abstrak

Dengan kemajuan teknologi kecerdasan buatan (AI) dan pembelajaran mesin, teknologi pengenalan suara akan terus berkembang. Dalam penelitian ini, akan dilakukan penelitian SLR (*Systematic Literature Review*) tentang pengenalan suara untuk mencari pembahasan tentang metode yang dipakai, bahasa yang diuji serta hambatan dan tantangan yang sering dihadapi. Sebanyak 2.400 artikel dikumpulkan dari 2 sumber data elektronik yang bersumber dari Scopus dan Semantic Scholar serta dalam rentang tahun 2020-2024, lalu disaring karena adanya duplikasi, tahapan kriteria inklusi dan eksklusi dan tahapan penilaian kualitas, sehingga didapat sekitar 32 artikel yang dipakai untuk menjawab tiga pertanyaan penelitian yang sudah dirumuskan melalui PICOC. Hasil yang didapat adalah terdapat 25 metode yang ditemukan dan metode CNN yang paling banyak dibahas dalam 6 artikel. Dari 28 bahasa yang ditemukan, bahasa Inggris merupakan bahasa yang paling banyak diuji dalam 7 artikel. Selain itu, terdapat 23 macam tantangan dan hambatan, yang paling banyak ditemui pada 17 artikel adalah sumber daya bahasa yang sedikit, dikarenakan hanya satu bahasa resmi dalam suatu negara dan ada pula yang hampir mengalami kepunahan sehingga tidak banyak tersedia untuk umum. Gangguan berupa kebisingan atau *noise* juga mengganggu dalam menyelesaikan penelitian tersebut. Lalu agar mencapai keakuratan yang tinggi dalam pengenalan suara, dibutuhkan data pelatihan yang besar. Penelitian SLR ini dapat mengidentifikasi tren dan metode terbaik dan terbukti efektif yang dapat dipakai di perangkat IoT, aplikasi *smartphone*, dan layanan *cloud*.

I. PENDAHULUAN

Kehidupan manusia tidak lepas dari perangkat teknologi, hal ini bisa kita lihat dari lingkungan sekitar karena segala pekerjaan dimudahkan oleh teknologi. Salah satu yang sering dipakai adalah pengenalan suara [1]. Pengenalan suara, juga dikenal sebagai pengenalan ucapan atau *speech recognition*, merujuk pada kemampuan teknologi untuk mengidentifikasi dan memahami ucapan manusia. Ini adalah cabang dari pengolahan bahasa alami yang bertujuan untuk mengonversi sinyal suara menjadi teks atau perintah yang dapat diinterpretasi oleh komputer. Kemampuan untuk berkomunikasi dengan perangkat melalui suara telah menjadi komponen kunci dalam perkembangan teknologi yang semakin canggih.

ASR (*Automatic Speech Recognition*) adalah teknologi dalam sebuah komputer untuk mengenali, memahami, dan mengonversi suara manusia menjadi teks secara otomatis [2]. ASR merupakan inti dari banyak aplikasi berbasis suara, seperti asisten virtual, sistem perintah suara, interaksi mesin manusia berupa chatbot berbasis suara [3], layanan transkripsi, dan layanan untuk penyandang disabilitas fisik untuk berkomunikasi atau mengontrol perangkat [4]. Cara kerja ASR adalah menangkap suara manusia melalui mikrofon sebagai inputan suara, sinyal suara diubah menjadi data digital dan dibersihkan dari gangguan seperti noise untuk meningkatkan suara, menganalisis fitur penting suara seperti frekuensi, durasi dan pola suara, lalu algoritma ASR membandingkan pola suara dengan model akustik dan bahasa yang telah dilatih sebelumnya sebagai pengenalan pola, dan konversi pola suara yang dikenali diubah menjadi teks. Terdapat tiga komponen utama ASR yaitu model akustik yaitu menghubungkan suara dengan fonem atau unit terkecil dari ucapan, model bahasa merupakan sebuah prediksi

* Corresponding author

ISSN 2622-2728 (online) 2622-271X (print)
doi: 10.32877/bt.v7i2.1888

kemungkinan urutan kata berdasarkan konteks untuk meningkatkan akurasi, dan decoder adalah hasil penggabungan dari model akustik dan model bahasa untuk menghasilkan teks akhir.

Pengenalan suara dengan AI (*Artificial Intelligence*) adalah penerapan teknologi kecerdasan buatan untuk meningkatkan kemampuan sistem pengenalan suara, sehingga dapat memahami, mengenali, dan menafsirkan ucapan manusia dengan lebih akurat dan efisien [5]. Dengan bantuan AI, pengenalan suara menjadi lebih pintar, lebih cerdas, fleksibel, dan memungkinkan interaksi manusia-mesin yang lebih alami. Seiring dengan kemajuan dalam bidang teknologi dan kecerdasan buatan, pengenalan suara telah mengalami peningkatan yang signifikan dalam akurasi dan efisiensi. Proses ini melibatkan analisis kompleks terhadap pola suara, termasuk pengenalan fonetik dan sintaktik, dengan memanfaatkan algoritma pembelajaran mesin, seperti *deep learning*, untuk meningkatkan performa sistem [6]. Meskipun masih ada tantangan seperti variasi aksen, kebisingan latar belakang, dan variasi gaya bicara seperti dialek, terus berkembangnya teknologi ini memberikan kontribusi besar terhadap peningkatan interaksi antara manusia dan mesin.

Pengenalan suara adalah teknologi penting yang memungkinkan komunikasi manusia dengan mesin menjadi lebih alami. Pengenalan suara tidak hanya membuka peluang baru dalam hal kenyamanan pengguna, tetapi juga meningkatkan aksesibilitas bagi individu dengan kebutuhan khusus [4]. Dengan terus meningkatnya daya komputasi dan pemahaman terhadap kompleksitas bahasa manusia, pengenalan suara terus menjadi bidang penelitian dan pengembangan yang menarik untuk menciptakan solusi yang lebih canggih dan terhubung dengan kehidupan sehari-hari manusia. Di masa depan pengenalan suara dengan AI bisa dalam penggunaan multi bahasa, dimana model AI akan semakin ahli dalam mengenali dan memahami berbagai bahasa dan dialek, peningkatan kemampuan pengenalan suara secara *real-time*, AI memahami emosi dan intonasi suara dalam interaksi yang lebih alami, dan integrasi lebih luar ke berbagai bidang seperti pendidikan, kesehatan, dan layanan pelanggan.

Penelitian SLR ini akan memberikan gambaran secara menyeluruh dan komprehensif. Dimulai dari mengumpulkan, menganalisis, dan mensintesis semua penelitian yang relevan dari berbagai sumber data elektronik terpercaya. Hal ini juga memberikan gambaran besar tentang apa yang telah diteliti dan apa yang masih menjadi tantangan. Dengan menganalisis studi-studi terdahulu, peneliti lain dapat mengidentifikasi tren, celah, dan metode terbaik. Dapat meningkatkan kualitas penelitian baru dengan menggunakan model dan metode yang telah terbukti efektif. Dan dapat membantu industri memilih pendekatan terbaik untuk membangun sistem pengenalan suara di perangkat IoT, aplikasi *smartphone*, dan layanan berbasis *cloud*.

II. TINJAUAN PUSTAKA

A. *Speech Recognition*

Pengenalan suara atau *speech recognition* adalah sebuah teknologi dalam komputer atau sistem digital dalam mengidentifikasi, memahami, dan memproses ucapan manusia ke dalam bentuk teks atau perintah yang dapat dipahami mesin [1], [2]. Teknologi ini bertujuan untuk mengubah suara analog menjadi teks atau instruksi digital. Contoh penerapan *speech recognition* bisa ditemukan dalam kehidupan sehari-hari, seperti asisten virtual contoh Siri, *Google Assistant*, dan Alexa yang ada pada *smartphone* atau perangkat lainnya. Lalu transkrip otomatis yang mengubah sebuah wawancara atau pidato menjadi teks. Penggunaan dalam navigasi yang memudahkan pengguna memberikan perintah suara pada aplikasi map atau peta. Membantu orang yang disabilitas untuk mengetik atau berkomunikasi. Dan berupa *chatbot* berbasis suara untuk melayani pelanggan.

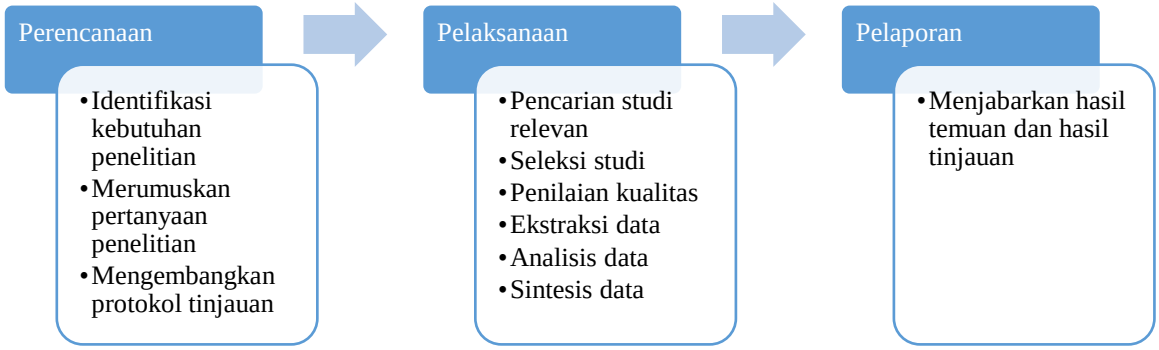
Tahapan dalam *speech recognition* yaitu *pre-processing*, *feature extraction*, *modeling* dan *decoding*. *Pre-processing* berupa membersihkan suara dari gangguan atau *noise* dan mengisolasi sinyal penting. *Feature extraction* adalah mengidentifikasi fitur penting dalam suara seperti intonasi, durasi, atau frekuensi. *Modeling* yaitu menggunakan algoritma dan model statistik misal dengan *deep learning* untuk mencocokkan suara dengan data *training*. *Decoding* merupakan kegiatan menerjemahkan pola suara yang dikenali menjadi teks atau perintah.

Hal-hal yang dibutuhkan dalam *speech recognition* adalah inputan suara pengguna yang diterima melalui perangkat seperti mikrofon. Lalu pengolahan sinyal dimana gelombang suara diubah menjadi data digital melalui proses seperti analisis frekuensi dan spektrum. Adanya pengenalan pola data suara dibandingkan dengan model yang telah dilatih sebelumnya untuk mengenali kata, frasa, atau pola suara. Dan output yang dihasilkan bisa berupa teks, atau perintah tertentu berdasarkan suara yang dikenali. Dengan perkembangan kecerdasan buatan (AI) [5], [6] dan pembelajaran mesin, teknologi pengenalan suara akan terus berkembang.

III. METODE

Tinjauan Literatur Sistematis atau SLR adalah sebuah metode pendekatan sistematis dan terstruktur untuk menemukan, mengevaluasi, dan menganalisis literatur atau penelitian yang berkaitan dengan topik atau pertanyaan tertentu [7]. Tujuan utama SLR adalah memberikan pemahaman yang komprehensif dan terpercaya mengenai temuan-temuan yang sudah ada serta mengindentifikasi kesenjangan penelitian untuk masa depan. SLR diawali dengan mendefinisikan protokol *review* yang mencakup pertanyaan penelitian atau tujuan *review*. SLR juga dilakukan berdasarkan kriteria pencarian yang telah ditetapkan untuk mendapatkan sebanyak-banyaknya pustaka yang relevan. Kemudian mendokumentasikan (secara eksplisit) kriteria dan strategi pencarian agar proses

dapat diulangi/direplikasi oleh peneliti lain. SLR mengharuskan adanya kriteria inklusi dan eksklusi yang jelas. Lalu menentukan secara jelas aspek informasi apa yang digali dari setiap pustaka yang di-*review*. SLR juga mensyaratkan sebuah *quantitative meta-analysis*. Metodologi dalam penelitian SLR ini dapat dilihat pada Gambar 1, yaitu sebagai berikut.



Gambar 1 Tahapan Penelitian

A. Tahap Perencanaan

Pada tahapan ini akan dilakukan identifikasi kebutuhan penelitian, merumuskan pertanyaan penelitian, dan mengembangkan protokol tinjauan. PICOC adalah suatu kerangka kerja yang digunakan dalam penelitian, untuk membantu merumuskan pertanyaan penelitian secara sistematis [11]. Penggunaan PICOC bertujuan agar penyusunan *Research Question* (RQ) menjadi lebih fokus dan tidak keluar dari ruang lingkup. PICOC merupakan singkatan dari *Population* (P), *Intervention* (I), *Comparison* (C), *Outcome* (O), dan *Context* (C). *Population* adalah target investigasi (seperti *software*, orang, instansi, dan lain-lain). *Intervention* adalah aspek yang akan diteliti (diinvestigasi) atau topik permasalahan yang akan diteliti. *Comparison* adalah aspek yang akan diperbandingkan. *Outcome* adalah luaran atau ukuran keberhasilan dari aspek yang diteliti. *Context* adalah konteks atau ruang lingkup yang akan diteliti (diinvestigasi). Pertanyaan penelitian dirumuskan melalui PICOC, yaitu *population* adalah populasi yang menjadi target pengumpulan data mengacu pada jurnal publikasi mengenai pengenalan suara. Lalu *intervension* mengenai penggunaan teknologi *machine learning* dan *deep learning* dalam pengenalan suara. Kemudian *comparison* yang mencakup analisis perbandingan tiap metode yang telah dilakukan. *Outcome* merupakan temuan penelitian ini mendapatkan akurasi yang paling tinggi. Dan terakhir *context* tentang penggunaan teknologi pengenalan suara pada *smartphone* dan perangkat lainnya.

Berdasarkan PICOC, dapat dirumuskan sebanyak tiga RQ atau pertanyaan penelitian. Pertanyaan penelitian yang pertama atau RQ1 adalah metode apa yang dipakai dalam mengenali suara manusia? Pertanyaan penelitian yang kedua atau RQ2 adalah terhadap bahasa apa saja yang diuji dalam penelitian? Lalu pertanyaan penelitian yang ketiga atau RQ3 adalah apa tantangan dan hambatan yang dihadapi para peneliti?

B. Tahap Pelaksanaan

Sebelum memulai penelitian, penting untuk menetapkan dan memvalidasi tahapan pencarian studi relevan untuk menjawab pertanyaan penelitian. Langkah ini melibatkan penilaian menyeluruh terhadap pilihan kata kunci dan terminologi yang akan digunakan dalam pencarian literatur serta pemilihan sumber basis data yang relevan. Tahapan pencarian studi relevan dirancang khusus yang disesuaikan dengan kebutuhan pertanyaan penelitian yang diajukan. Setelah strategi pencarian dikembangkan, pencarian yang luas dan komprehensif dilakukan melalui sumber elektronik yang signifikan.

Penelitian ini melibatkan 2 database akademis utama yaitu Scopus dan Semantic Scholar dimana pencarian sumber data elektronik ini menggunakan *software Publish or Perish*. Dengan mengikuti metodologi yang telah ditetapkan, penelitian ini bertujuan untuk mengumpulkan sebanyak mungkin studi yang relevan untuk menjawab pertanyaan penelitian. Pemilihan database akademis Scopus dikarenakan mempunyai standar kualitas yang tinggi. Dan untuk Semantic Scholar dilakukan karena sebagai alat pencarian gratis dan didukung oleh AI untuk membantu menemukan artikel yang relevan. Tabel 1 menunjukkan sumber data elektronik dengan URL-nya.

TABEL 1
SUMBER DATA

Sumber Data	URL
Scopus	www.scopus.com
Semantic Scholar	www.semanticscholar.org

Setelah mengidentifikasi basis data yang sesuai, langkah berikutnya adalah menyusun berbagai kata kunci dan kombinasinya. Kata kunci berikut digunakan dalam penelitian ini yaitu *speech*, *speech recognition*, *deep learning*, dan *machine learning*. Pencarian kata kunci tingkat lanjut menggunakan formulasi dengan operator Boolean seperti DAN dan ATAU. Kombinasi kata kunci yang digunakan dalam penelitian yaitu "*Speech Recognition Machine Learning*" ATAU "*Speech Recognition Deep Learning*".

Selanjutnya adalah tahapan kriteria seleksi studi untuk meninjau ketentuan dalam memilih artikel. Kriteria seleksi studi ini akan menentukan apakah artikel tersebut layak digunakan sebagai sumber penelitian atau tidak. Kriteria pertama adalah kriteria inklusi yang dalam penelitian ini yaitu jenis artikel merupakan jurnal artikel maupun *conference paper*, artikel harus ditulis dalam bahasa Inggris dan dipublikasikan dari tahun 2020 hingga 2024 dan terbuka untuk umum (*Open Access*). Artikel harus mempunyai relasi dengan bidang utama yang menjadi fokus penelitian yaitu *computer science*, *speech communication*, dan kecerdasan buatan atau *artificial intelligence*. Penetapan kriteria inklusi ini bertujuan untuk mengumpulkan data yang komprehensif dan terkini tentang pengembangan dan penerapan teknologi dalam pengenalan suara serta interaksinya dengan kecerdasan buatan dalam menciptakan solusi inovatif untuk tantangan masa kini dan masa mendatang. Dengan membatasi cakupan publikasi artikel hingga 5 tahun terakhir, penelitian SLR dapat fokus pada pendekatan modern yang lebih efisien, akurat, dan relevan dibandingkan metode lama. Kriteria kedua adalah kriteria ekslusi yang dalam penelitian ini adalah artikel *review* seperti dalam bentuk tinjauan studi literatur atau SLR dan berupa buku, serta artikel yang ditulis selain bahasa Inggris tidak akan diikutsertakan dalam penelitian ini. Penjabaran syarat atau kriteria inklusi dan kriteris eksklusi tertuang pada Tabel 2.

TABEL 2
KRITERIA INKLUSI DAN EKSKLUSI

Kriteria	Syarat
Inklusi	- Artikel dan *conference paper* - Dalam Bahasa Inggris - Dipublikasi dari tahun 2020-2024 - *Open access* - Bidang *computer science* - Bidang *speech communication* - Bidang *artificial intelligence*
Eksklusi	- Artikel review dan buku - Tidak ditulis dalam Bahasa Inggris

Sebanyak 2.400 artikel terkumpul dengan masing-masing sumber data *Scopus* yaitu 400 dan *Semantic Scholar* sejumlah 2.000 artikel. Dari ke 2.400 artikel tersebut dikurangi oleh jurnal yang terduplikasi sebanyak 510 artikel, sehingga total artikel yang didapat adalah 1.890 artikel. Kemudian 1.890 artikel disaring berdasarkan selain tipe artikel dan *conference paper* sebanyak 195 artikel. Disaring kembali pada ketepatan kata kunci dalam judul dan abstrak sejumlah 1.245 artikel, tidak ditulis dalam Bahasa Inggris sebanyak 2 artikel, dan tidak sesuai bidang *computer science*, *speech communication* dan *artificial intelligence* dengan jumlah 310 artikel. Total artikel yang didapat untuk ke tahap berikutnya adalah sebanyak 138 artikel.

Kemudian tahap penilaian kualitas (*Quality Assesment*) untuk memastikan keakuratan dalam kesimpulan penelitian SLR ini. Kategori jawaban dibagi menjadi 3 yaitu Ya, Tidak, dan Sebagian. Kategori jawaban "Ya" jika penelitian tersebut menunjukkan penjelasan komprehensif dan jelas, dan mendapatkan skor nilai 1. Kategori jawaban "Tidak" jika penjelasan tidak ada dan diberi skor nilai 0. Kategori jawaban "Sebagian" jika penjelasan tidak dijelaskan dengan jelas atau hanya sebagian dan diberi skor nilai 0,5. Berikut pada Tabel 3 adalah daftar pertanyaan yang berbentuk survei sebagai penilaian kualitas.

TABEL 3
DAFTAR PERTANYAAN DALAM PENILAIAN KUALITAS

No	Pertanyaan	Jawaban
1	Apakah artikel disitasi oleh peneliti lain?	Ya/Tidak
2	Apakah tujuan penelitian dan pertanyaan penelitian dijabarkan dengan jelas?	Ya/Tidak/Sebagian
3	Apakah metode *speech recognition* pada penelitian dijelaskan?	Ya/Tidak/Sebagian
4	Apakah bahasa manusia yang diuji dengan metode tersebut?	Ya/Tidak/Sebagian
5	Apakah tantangan dan hambatan yang ditemukan peneliti disebutkan dengan jelas?	Ya/Tidak/Sebagian

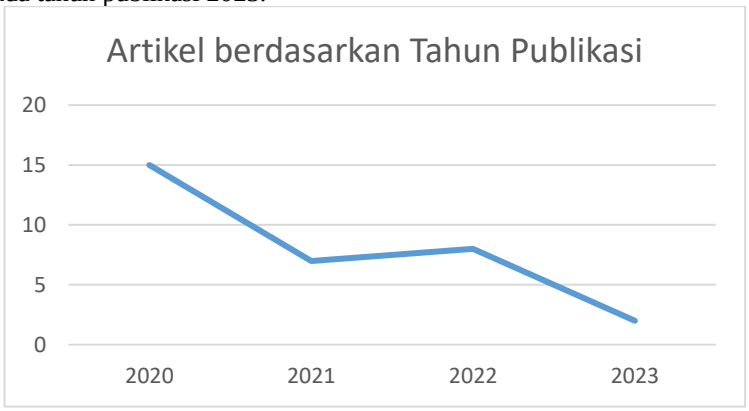
Nilai sebuah artikel akan ditentukan berdasarkan jawaban dari pertanyaan-pertanyaan tersebut. Skor tertinggi yang dapat dicapai adalah 5, sedangkan skor terendah adalah 0. Berdasarkan penilaian tersebut, ambang batas skor minimum untuk artikel yang akan dimasukkan ke dalam penelitian ini adalah sebesar 4. Untuk artikel yang memiliki skor di angka 4 atau lebih menandakan artikel tersebut menandakan pemahaman yang komprehensif dan memiliki tinjauan yang berkualitas tinggi dalam kontribusi untuk pembahasan selanjutnya.

Dari 138 artikel yang tersedia dan telah melalui tahap penilaian kualitas, hanya 32 artikel yang dapat diikutsertakan dalam pembahasan dengan rincian skor di angka 4 sebanyak 31 artikel dan skor di angka 4,5 sebanyak 1 artikel. Sisanya 105 jurnal memiliki skor di bawah 4 dengan rincian berikut. Skor di angka 2,5 sebanyak 1 artikel, skor di angka 3 sebanyak 20 artikel, skor di angka 3,5 sebanyak 85 artikel. Rincian skor nilai dan jumlah artikel dalam tahap penilaian kualitas ditunjukkan pada Tabel 4.

TABEL 4
SKOR NILAI DAN JUMLAH ARTIKEL TAHAP PENILAIAN KUALITAS

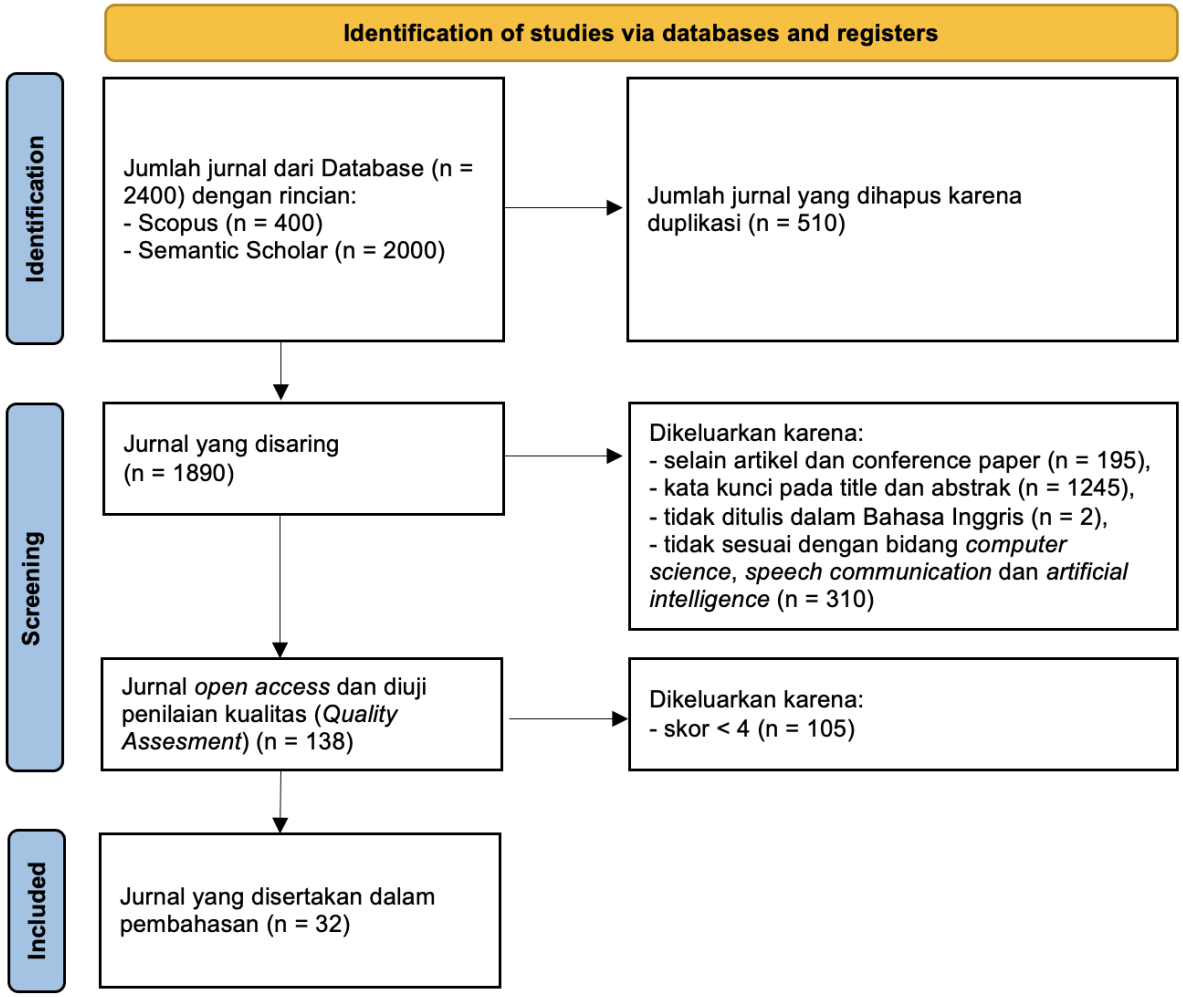
No	Skor Nilai	Jumlah Artikel
1	2,5	1
2	3	20
3	3,5	85
4	4	31
5	4,5	1

Gambar 2 menyatakan penyebaran tahun publikasi dari ke-32 artikel tersebut. Tahun publikasi 2020 terdapat 15 artikel, diikuti tahun 2021 sebanyak 7 artikel, tahun 2022 sebanyak 8 artikel dan tahun 2023 sebanyak 2 artikel. Penyebaran ini dapat dilihat lebih banyak pada tahun publikasi 2020 setelah melalui tahapan penilaian kualitas, dan semakin sedikit pada tahun publikasi 2023.



Gambar 2 Jumlah Artikel per Tahun Publikasi Setelah Tahapan Penilaian Kualitas

PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-analyses*) sebuah metodologi yang dirancang untuk membantu peneliti melaporkan hasil review dan meta-analisis sistematis dengan cara yang mudah dipahami dan lengkap [14]. PRISMA bertujuan untuk meningkatkan kualitas pelaporan dengan memberikan daftar elemen penting yang harus disertakan dalam sebuah laporan penelitian sistematis. PRISMA *Flow Diagram* memberikan visualisasi alur proses seleksi studi yang terdiri dari identifikasi jumlah studi yang ditemukan dari pencarian database, penyaringan mengevaluasi berdasarkan kriteria inklusi dan eksklusi, kelayakan studi berdasarkan setelah dibaca, lalu *included* atau penyertaan studi yang akan disertakan dalam analisis. Dimulai dari tahap identifikasi, *screening* dan *included* yang menghasilkan 32 artikel yang akan dibahas. Kesimpulan tahap pelaksanaan dibuat dalam sebuah kerangka kerja PRISMA yang dapat dilihat pada Gambar 3.



Gambar 3 Kerangka kerja PRISMA dalam seleksi artikel

IV. HASIL

Pada tahap hasil dilakukan analisa untuk menjawab tiap RQ atau pertanyaan penelitian dari ke-32 artikel tersebut.

A. RQ1: Metode apa yang dipakai dalam mengenali suara manusia?

Pertanyaan penelitian 1 ini mengarah pada tinjauan komprehensif dari artikel yang ada untuk mengidentifikasi elemen-elemen yang secara terus menerus muncul dalam teknologi pengenalan suara. Hasil tinjauan studi yang komprehensif mengungkapkan pendekatan atau metode penting dalam membentuk teknologi ini. Terdapat 25 metode yang ditemukan, dipakai dan dibahas penggunaannya dalam ke-32 artikel tersebut.

Yang pertama adalah metode DNN (*Deep Neural Networks*) yang menggunakan jaringan saraf dalam untuk memetakan fitur suara ke teks atau fonem: [16], [17], [18]. CNN (*Convolutional Neural Networks*) menganalisis fitur suara yang telah dikonversi menjadi representasi visual atau spectrogram: [19], [20], [21], [22], [23], [24]. Wav2Vec adalah model berbasis *deep learning* yang dirancang untuk pengenalan suara: [25], [26], [27], [28]. CNN-LSTM adalah gabungan model CNN dan LSTM (*Long Short-Term Memory*): [29], [30]. ResNet (*Residual Network*) adalah jenis arsitektur dari CNN: [21]. BiLSTM (*Bidirectional* LSTM) adalah pengembangan dari LSTM: [21], [31], [32]. End-to-End-Transformer adalah gabungan model End-to-End dan Transformer: [33]. DNN-HMM merupakan gabungan model DNN dan HMM (*Hidden Markov Model*): [33]. TDNN (*Time-Delay Neural Network*) adalah jenis jaringan saraf tiruan yang dirancang untuk bekerja dengan data sekuensial, seperti sinyal audio atau teks: [34], [18]. RNN (*Recurrent Neural Networks*) memproses data sekuensial dengan mempertimbangkan informasi dari langkah-langkah sebelumnya: [35], [36], [37]. CTC (*Connectionist Temporal Classification*) yaitu algoritma *loss* untuk pelatihan model *sequence-to-sequence*, memungkinkan pengenalan suara tanpa pelabelan waktu yang presisi: [38], [39], [40], [41]. Transformer menggunakan mekanisme attention untuk memodelkan hubungan global antara elemen dalam data sekuensial: [36], [42], [31], [43]. CNN–RBM–ASAT adalah gabungan dari model CNN, RBM (*Restricted Boltzmann Machine*) dan ASAT: [44]. GMM-HMM

adalah gabungan model GMM (*Gaussian Mixture Models*) dan HMM: [18]. TDNN-LSTM yaitu gabungan model TDNN dan LSTM: [18]. Conformer merupakan gabungan CNN dan Transformer: [31]. HMM menggunakan probabilitas untuk memodelkan hubungan antara urutan suara (speech signals) dan kata-kata atau fonem (unit suara terkecil dalam bahasa): [32]. LSTM adalah model khusus RNN yang mampu menangani dependensi jangka panjang dalam data sekuensial: [32]. LSTM-DBN yaitu gabungan model LSTM dan DBN (*Deep Belief Network*): [32]. DLSTM (*Distributed* LSTM) merupakan pengembangan dari LSTM: [32]. DBLSTM (*Deep Bidirectional* LSTM) yaitu pengembangan dari LSTM: [32]. DBLSTM-DNN adalah gabungan model DBLSTM dan DNN: [32]. BERT (*Bidirectional Encoder Representations from Transformers*) adalah model *deep learning* berbasis Transformer yang dirancang untuk memahami konteks kata dalam sebuah teks secara *bidirectional*: [45]. RNN-CTC adalah gabungan model RNN dan CTC: [46]. DBN adalah jenis DNN yang terdiri dari beberapa lapisan berbasis probabilistik: [47]. Tabel 5 menunjukkan 25 metode yang ditemukan dan artikel mana saja yang membahas metode tersebut.

TABEL 5
METODE PENGENALAN SUARA

No	Metode	Artikel
1	DNN	[16], [17], [18]
2	CNN	[19], [20], [21], [22], [23], [24]
3	Wav2vec2	[25], [26], [27], [28]
4	CNN-LSTM	[29], [30]
5	ResNet	[21]
6	BiLSTM	[21], [31], [32]
7	End-to-End-Transformer	[33]
8	DNN-HMM	[33]
9	TDNN	[34], [18]
10	RNN	[35], [36], [37]
11	CTC	[38], [39], [40], [41]
12	Transformer	[36], [42], [31], [43]
13	CNN–RBM–ASAT	[44]
14	GMM-HMM	[18]
15	TDNN-LSTM	[18]
16	Conformer	[31]
17	HMM	[32]
18	LSTM	[32]
19	LSTM-DBN	[32]
20	DLSTM	[32]
21	DBLSTM	[32]
22	DBLSTM-DNN	[32]
23	BERT	[45]
24	RNN-CTC	[46]
25	DBN	[47]

B. RQ2: Terhadap bahasa apa saja yang diuji dalam penelitian?

Pertanyaan penelitian 2 ini menjabarkan tentang bahasa manusia yang diuji dalam hal akurasi dan efektivitas dalam teknologi pengenalan suara. Terdapat 28 bahasa yang ditemukan dan diuji dalam ke-32 artikel tersebut. Diantaranya adalah bahasa Inggris yang merupakan bahasa resmi dunia pertama dan paling banyak dipakai di seluruh dunia: [19], [25], [35], [38], [44], [46], [43]. Bahasa Arab yaitu bahasa resmi dunia yang ketiga: [19], [29], [30]. Bahasa Amazigh atau bahasa Berber yang merupakan cabang dari rumpun bahasa Afro-Asia: [19]. Bahasa Vietnam adalah bahasa nasional dan resmi di Vietnam: [20]. Bahasa Portugis adalah bahasa resmi di Portugal dan Brazil: [17], [26], [41]. Bahasa Nepal adalah bahasa resmi di Nepal: [21]. Bahasa Uzbek yaitu bahasa resmi di Uzbekistan dan termasuk dalam rumpun bahasa Turkik: [33]. Bahasa Belanda adalah bahasa resmi di Belanda: [34], [41]. Bahasa Tagalog yaitu bahasa resmi di Filipina: [22]. Bahasa Dari yaitu bahasa resmi di Afganistan: [23]. Bahasa Perancis yang adalah bahasa resmi dunia dan bahasa Italia yaitu bahasa resmi di Italia: [38]. Bahasa Mandarin merupakan bahasa resmi dunia yang kedua: [39], [41]. Bahasa Indonesia adalah bahasa resmi di Indonesia: [36]. Bahasa Japhug termasuk rumpun bahasa Trans-Himalaya (Sino-Tibet): [42]. Bahasa Seneca adalah bahasa yang dipakai oleh Suku Seneca di Amerika Utara: [24]. Bahasa Turki yang merupakan bahasa resmi di Turki dan Siprus Utara: [40], [37]. Bahasa Mongol yaitu bahasa resmi di Mongolia, dan bahasa Haiti yang adalah bahasa resmi di Haiti: [41]. Bahasa Rumania yaitu bahasa resmi di Rumania: [37]. Bahasa Jerman yang merupakan bahasa pengantar di kawasan Eropa Tengah: [37]. Bahasa Sinhala adalah bahasa resmi di Sri Lanka: [18]. Bahasa Telugu termasuk rumpun bahasa Dravida di India bagian selatan, bahasa Tamil yang juga termasuk rumpun bahasa Dravida di Tamil Nadu di negara India, dan Bahasa Gujarati yang merupakan bahasa resmi negara bagian Gujarat di negara India: [31], [28]. Bahasa Persia adalah salah satu rumpun bahasa Iran: [32]. Bahasa Bengali adalah bahasa resmi di Bangladesh: [45]. Bahasa Thai yaitu bahasa resmi di Thailand: [27]. Tabel 6 menunjukkan 28 bahasa yang dipakai dan diuji berikut artikel mana saja yang membahas bahasa-bahasa tersebut.

TABEL 6

BAHASA YANG DIUJI

No	Bahasa	Artikel
1	Inggris	[19], [25], [35], [38], [44], [46], [43]
2	Arab	[19], [29], [30]
3	Amazigh	[19]
4	Vietnam	[20]
5	Portugis	[17], [26], [41]
6	Nepal	[21]
7	Uzbek	[33]
8	Belanda	[34], [41]
9	Tagalog	[22]
10	Dari	[23]
11	Perancis	[38]
12	Italia	[38]
13	Mandarin	[39], [41]
14	Indonesia	[36]
15	Japhug	[42]
16	Seneca	[24]
17	Turki	[40], [37]
18	Mongol	[41]
19	Haiti	[41]
20	Rumania	[37]
21	Jerman	[37]
22	Sinhala	[18]
23	Telugu	[31], [28]
24	Tamil	[31], [28]
25	Gujarati	[31], [28]
26	Persia	[32]
27	Bengali	[45]
28	Thai	[27]

C. RQ3: Apa tantangan dan hambatan yang dihadapi para peneliti?

Pertanyaan penelitian ke-3 menjawab tantangan dan hambatan peneliti yang ditemukan berdasarkan ke-32 artikel tersebut. Terdapat sejumlah 23 tantangan dan hambatan yang dihadapi oleh peneliti. Yang pertama adalah percakapan langsung secara *real-time* tanpa waktu jeda: [16]. Lalu membutuhkan data pelatihan dalam volume besar: [43], [17], [33], [36]. Noise atau gangguan kebisingan: [19], [17], [23], [36]. Tidak bisa mengenali banyak atau multi-bahasa saat melakukan pengenalan suara: [19], [31]. Sumber daya bahasa dalam jumlah sedikit yang tersedia untuk umum: [20], [25], [17], [33], [34], [26], [22], [23], [36], [42], [24], [40], [18], [31], [45], [27], [28]. Kompleksitas morfologi yang merupakan perubahan bentuk kata terhadap arti kata: [29]. Kosakata yang banyak dan luas: [29], [18]. Diakritik adalah tanda baca tambahan yang mengubah nilai fonetis huruf: [29]. Gap senyap atau ada jeda waktu dalam waktu lama tanpa suara: [21]. Overfitting atau model tidak dapat membuat prediksi dan terlalu cocok dengan data pelatihan: [30], [41]. Aksen tiap individu yang berbeda-beda: [35]. Distilasi pengetahuan yaitu teknik melatih model siswa (model kecil) dari model guru (model besar): [38]. Bahasa kedua yang berarti bukan bahasa ibu di negara tersebut dan hanya sebagai pembelajaran: [39]. Bahasa Inggris pada sistem yang sudah ditanam pada teknologi: [44]. Fonem yang merupakan satuan bunyi terkecil: [37], [31]. Akurasi dalam mengenali kata dan makna: [32], [46], [27]. Efisiensi waktu dalam memproses: [32], [46]. Tanda baca memberikan petunjuk mengenai intonasi, penekanan dan artikulasi: [45]. Nada suara atau tinggi rendahnya bunyi, kecepatan bicara seseorang berbeda-beda tiap individu, ritme dan intonasi: [47]. Sumber ASR tertutup yang berarti tidak *open-source* untuk umum: [27]. Tabel 7 menunjukkan 23 tantangan dan hambatan berikut dengan artikel mana saja yang membahas tantangan dan hambatan tersebut.

TABEL 7
TANTANGAN DAN HAMBATAN

No	Metode	Artikel
1	Percakapan langsung	[16]
2	Data pelatihan yang besar	[43], [17], [33], [36]
3	Noise	[19], [17], [23], [36]
4	Multi-bahasa	[19], [31]
5	Sumber daya sedikit	[20], [25], [17], [33], [34], [26], [22], [23], [36], [42], [24], [40], [18], [31], [45], [27], [28]
6	Kompleksitas morfologi	[29]
7	Kosakata luas	[29], [18]
8	Diakritik	[29]
9	Gap senyap	[21]
10	Overfitting	[30], [41]
11	Aksen	[35]
12	Distilasi pengetahuan	[38]
13	Bahasa kedua	[39]

14	Bahasa Inggris pada sistem	[44]
15	Fonem	[37], [31]
16	Akurasi	[32], [46], [27]
17	Efisiensi	[32], [46]
18	Tanda baca	[45]
19	Nada suara	[47]
20	Kecepatan bicara	[47]
21	Ritme	[47]
22	Intonasi	[47]
23	Sumber ASR tertutup	[27]

V. PEMBAHASAN

Berikut akan membahas tentang hasil analisa dari tiga pertanyaan penelitian secara mendetail. Pertanyaan penelitian pertama (RQ1) tentang metode apa yang dipakai dalam mengenali suara. Dari 25 metode yang ditemukan, metode yang paling banyak ditemukan dari ke-32 artikel tersebut adalah CNN sebanyak 6 artikel, diikuti oleh metode *Wav2vec2*, CTC dan *Transformer*. Ada juga metode yang menggabungkan dua atau lebih model dengan harapan bisa meningkatkan akurasi pengenalan suara, seperti CNN-LSTM, End-to-End-Transformer, DNN-HMM, CNN-RBM-ASAT, GMM-HMM, TDNN-LSTM, Conformer, LSTM-DBN, DBLSTM-DNN, dan RNN-CTC. Pertanyaan penelitian kedua (RQ2) tentang terhadap bahasa apa saja yang diuji dalam penelitian. Dari 28 bahasa yang ditemukan, bahasa Inggris adalah bahasa yang paling banyak ditemukan dalam 7 artikel. Bahasa Inggris yang dipakai juga memiliki berbagai banyak aksen yang berbeda seperti Inggris-India, Inggris-Jerman, Inggris-Spanyol dan Inggris-Perancis. Berikutnya bahasa Arab dan bahasa Portugis yang juga banyak ditemukan dalam masing-masing sebanyak 3 artikel. Pertanyaan penelitian ketiga (RQ3) tentang tantangan dan hambatan yang dihadapi para peneliti. Terdapat 23 tantangan dan hambatan yang dihadapi oleh para peneliti dalam 32 artikel tersebut. Salah satu yang paling banyak dihadapi adalah sumber daya bahasa yang tidak banyak tersedia untuk umum dalam 17 artikel. Sumber daya yang dimaksud adalah bahasa yang hanya dipakai pada satu negara saja seperti bahasa Vietnam, bahasa Nepal, bahasa Uzbek, bahasa Belanda, bahasa Tagalog, bahasa Dari, bahasa Italia, bahasa Indonesia, bahasa Haiti, bahasa Rumania, bahasa Sinhala, bahasa Persia, bahasa Bengali dan bahasa Thai. Sumber daya bahasa yang tidak banyak juga dikarenakan bahasa tersebut hampir punah keberadaannya seperti bahasa Seneca dan bahasa Japhug. Tantangan dan hambatan selanjutnya adalah memerlukan data pelatihan yang besar sebanyak 4 artikel, hal ini dikarenakan untuk mencapai tingkat akurasi yang tinggi. Lalu tantangan dan hambatan berikutnya yang sering ditemui adalah adanya *noise* dari lingkungan sekitar yang mengganggu dalam keakuratan pengenalan suara sebanyak 4 artikel.

Dalam penelitian SLR ini memiliki keterbatasan dan kendala yang mempengaruhi dalam menjawab 3 pertanyaan penelitian. Pertama, penelitian SLR ini terbatas pada artikel jurnal dan *conference paper*, tidak termasuk jenis publikasi lainnya seperti *review*, buku dan survei yang mungkin dapat memberikan pengetahuan penting. Kedua, artikel yang dipilih hanya yang berbahasa Inggris, yang mungkin mengabaikan pengetahuan penting yang telah dipublikasikan dalam bahasa lain. Penelitian mendatang diharapkan dapat membantu peneliti lain mengembangkan metode lain untuk mengatasi tantangan dan hambatan yang sering ditemui dalam ke-32 artikel tersebut, men.

VI. KESIMPULAN

Hasil dari pertanyaan penelitian pertama adalah metode CNN yang paling banyak ditemukan dalam 6 artikel atau sekitar 13% dari 25 metode lainnya. Hasil dari pertanyaan penelitian kedua adalah bahasa Inggris yang paling banyak dipakai dalam 7 artikel atau sekitar 16% dari 28 bahasa lainnya. Hasil dari pertanyaan penelitian ketiga adalah sumber daya bahasa yang sedikit dianggap paling banyak dihadapi para peneliti sebagai tantangan dan hambatan dalam pengenalan suara yang ditemukan pada 17 artikel atau sekitar 33% dari 23 macam tantangan dan hambatan lainnya. Hasil penelitian SLR ini dapat digunakan untuk mencari metode terbaru dan efektif, meneliti pengenalan suara terhadap bahasa yang jarang diuji, dan dapat mengatasi berbagai tantangan dan hambatan yang akan dihadapi mendatang.

REFERENCES

[1] J. Meng, J. Zhang, and H. Zhao, "Overview of the Speech Recognition Technology," in *2012 Fourth International Conference on Computational and Information Sciences*, Chongqing, China: IEEE, Aug. 2012, pp. 199–202. doi: 10.1109/ICCIS.2012.202.

[2] B. H. Juang and L. R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development".

[3] A. Shenoy, S. Bodapati, and K. Kirchhoff, "ASR Adaptation for E-commerce Chatbots using Cross-Utterance Context and Multi-Task Language Modeling," in *Proceedings of The 4th Workshop on e-*

Commerce and NLP, Online: Association for Computational Linguistics, 2021, pp. 18–25. doi: 10.18653/v1/2021.ecnlp-1.3.

[4] J. Noyes and C. Frankish, "Speech recognition technology for individuals with disabilities," *Augmentative and Alternative Communication*, vol. 8, no. 4, pp. 297–303, Jan. 1992, doi: 10.1080/07434619212331276333.

[5] Z. Leini and S. Xiaolei, "Study on Speech Recognition Method of Artificial Intelligence Deep Learning," *J. Phys.: Conf. Ser.*, vol. 1754, no. 1, p. 012183, Feb. 2021, doi: 10.1088/1742-6596/1754/1/012183.

[6] Ubon Ratchathani Rajabhat University, Thailand and N. K. Dennis, "Using AI-Powered Speech Recognition Technology to Improve English Pronunciation and Speaking Skills," *ije*, vol. 12, no. 2, pp. 107–126, Aug. 2024, doi: 10.22492/ije.12.2.05.

[7] D. Gough, J. Thomas, and S. Oliver, "An introduction to systematic reviews," 2017.

[8] A. Dhouib, A. Othman, O. El Ghoul, M. K. Khribi, and A. Al Sinani, "Arabic Automatic Speech Recognition: A Systematic Literature Review," *Applied Sciences*, vol. 12, no. 17, p. 8898, Sep. 2022, doi: 10.3390/app12178898.

[9] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[10] V. Bhardwaj *et al.*, "Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review," *Applied Sciences*, vol. 12, no. 9, p. 4419, Apr. 2022, doi: 10.3390/app12094419.

[11] A. Booth, A. Sutton, and D. Papaioannou, *Systematic approaches to a successful literature review*, Second edition. Los Angeles: Sage, 2016.

[12] M. Bruzza, A. Cabrera, and M. Tupia, "Survey of the state of art based on PICOC about the use of artificial intelligence tools and expert systems to manage and generate tourist packages," in *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, Dubai: IEEE, Dec. 2017, pp. 290–296. doi: 10.1109/ICTUS.2017.8286021.

[13] W. Mengist, T. Soromessa, and G. Legese, "Method for conducting systematic literature review and meta-analysis for environmental science research," *MethodsX*, vol. 7, p. 100777, 2020, doi: 10.1016/j.mex.2019.100777.

[14] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *PLoS Med*, vol. 18, no. 3, p. e1003583, Mar. 2021, doi: 10.1371/journal.pmed.1003583.

[15] Y. Harie, B. P. Gautam, and K. Wasaki, "Computer Vision Techniques for Growth Prediction: A Prisma-Based Systematic Literature Review," *Applied Sciences*, vol. 13, no. 9, p. 5335, Apr. 2023, doi: 10.3390/app13095335.

[16] D. Jiang *et al.*, "A GDPR-compliant Ecosystem for Speech Recognition with Transfer, Federated, and Evolutionary Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 3, pp. 1–19, Jun. 2021, doi: 10.1145/3447687.

[17] I. Quintanilha, S. Netto, and L. Biscainho, "An open-source end-to-end ASR system for Brazilian Portuguese using DNNs built from newly assembled corpora," *JCIS*, vol. 35, no. 1, pp. 230–242, 2020, doi: 10.14209/jcis.2020.25.

[18] H. Karunathilaka, V. Welgama, T. Nadungodage, and R. Weerasinghe, "Low-resource Sinhala Speech Recognition using Deep Learning," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka: IEEE, Nov. 2020, pp. 196–201. doi: 10.1109/ICTer51097.2020.9325468.

[19] K. Choutri, M. Lagha, S. Meshoul, M. Batouche, Y. Kacel, and N. Mebarkia, "A Multi-Lingual Speech Recognition-Based Framework to Human-Drone Interaction," *Electronics*, vol. 11, no. 12, p. 1829, Jun. 2022, doi: 10.3390/electronics11121829.

[20] Q. H. Nguyen and T.-D. Cao, "A Novel Method for Recognizing Vietnamese Voice Commands on Smartphones with Support Vector Machine and Convolutional Neural Networks," *Wireless Communications and Mobile Computing*, vol. 2020, pp. 1–9, Mar. 2020, doi: 10.1155/2020/2312908.

[21] M. Dhakal, A. Chhetri, A. K. Gupta, P. Lamichhane, S. Pandey, and S. Shakya, "Automatic speech recognition for the Nepali language using CNN, bidirectional LSTM and ResNet," in *2022 International Conference on Inventive Computation Technologies (ICICT)*, Nepal: IEEE, Jul. 2022, pp. 515–521. doi: 10.1109/ICICT54344.2022.9850832.

[22] F. R. Jr. Arnel Fajardo, "Convolutional Neural Network for Automatic Speech Recognition of Filipino Language," *IJATCSE*, vol. 9, no. 1.1 S I, pp. 34–40, Feb. 2020, doi: 10.30534/ijatcse/2020/0791.12020.

[23] M. Dawodi, J. A. Baktash, T. Wada, N. Alam, and M. Z. Joya, "Dari Speech Classification Using Deep Convolutional Neural Network," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Vancouver, BC, Canada: IEEE, Sep. 2020, pp. 1–4. doi: 10.1109/IEMTRONICS51293.2020.9216370.

[24] R. Jimerson, R. Ptucha, and E. Prud'hommeaux, "Fully Convolutional ASR for Less-Resourced Endangered Languages".
[25] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A WAV2VEC2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition," *IEEE Access*, vol. 11, pp. 46938–46948, 2023, doi: 10.1109/ACCESS.2023.3275106.
[26] L. R. S. Gris, E. Casanova, F. S. de Oliveira, A. da S. Soares, and A. C. Junior, "Brazilian Portuguese Speech Recognition Using Wav2vec 2.0," Dec. 22, 2021, *arXiv*: arXiv:2107.11414. doi: 10.48550/arXiv.2107.11414.
[27] W. Phatthiyaphaibun, C. Chaksangchaichot, P. Limkonchotiwat, E. Chuangsuwanich, and S. Nutanong, "Thai Wav2Vec2.0 with CommonVoice V8," Aug. 09, 2022, *arXiv*: arXiv:2208.04799. doi: 10.48550/arXiv.2208.04799.
[28] K. D. N, P. Wang, and B. Bozza, "Using Large Self-Supervised Models for Low-Resource Speech Recognition," in *Interspeech 2021*, ISCA, Aug. 2021, pp. 2436–2440. doi: 10.21437/Interspeech.2021-631.
[29] H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, and Z. T. Fayed, "Arabic speech recognition using end-to-end deep learning," *IET Signal Processing*, vol. 15, no. 8, pp. 521–534, Oct. 2021, doi: 10.1049/sil2.12057.
[30] H. Alsayadi, A. Abdelhamid, I. Hegazy, and Z. Taha, "Data Augmentation for Arabic Speech Recognition Based on End-to-End Deep Learning," *IJICIS*, vol. 21, no. 2, pp. 50–64, Jul. 2021, doi: 10.21608/ijicis.2021.73581.1086.
[31] K. D. N, "Multilingual Speech Recognition for Low-Resource Indian Languages using Multi-Task conformer," Sep. 10, 2021, *arXiv*: arXiv:2109.03969. doi: 10.48550/arXiv.2109.03969.
[32] H. Veisi and A. Haji Mani, "Persian speech recognition using deep learning," *Int J Speech Technol*, vol. 23, no. 4, pp. 893–905, Dec. 2020, doi: 10.1007/s10772-020-09768-x.
[33] A. Mukhamadiyev, I. Khujayarov, O. Djuraev, and J. Cho, "Automatic Speech Recognition Method Based on Deep Learning Approaches for Uzbek Language," *Sensors*, vol. 22, no. 10, p. 3683, May 2022, doi: 10.3390/s22103683.
[34] P. Wang and H. Van Hamme, "Benefits of pre-trained mono- and cross-lingual speech representations for spoken language understanding of Dutch dysarthric speech," *J AUDIO SPEECH MUSIC PROC.*, vol. 2023, no. 1, p. 15, Apr. 2023, doi: 10.1186/s13636-023-00280-z.
[35] P. Dubey and B. Shah, "Deep Speech Based End-to-End Automated Speech Recognition (ASR) for Indian-English Accents".
[36] S. Suyanto, A. Arifianto, A. Sirwan, and A. P. Rizaendra, "End-to-End Speech Recognition Models for a Low-Resourced Indonesian Language," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, Yogyakarta, Indonesia: IEEE, Jun. 2020, pp. 1–6. doi: 10.1109/ICoICT49345.2020.9166346.
[37] C. Liu, F. Zhang, D. Le, S. Kim, Y. Saraf, and G. Zweig, "Improving RNN Transducer Based ASR with Auxiliary Tasks," Nov. 09, 2020, *arXiv*: arXiv:2011.03109. doi: 10.48550/arXiv.2011.03109.
[38] Y. Gao, T. Parcollet, and N. Lane, "Distilling Knowledge from Ensembles of Acoustic Models for Joint CTC-Attention End-to-End Speech Recognition," Jul. 04, 2021, *arXiv*: arXiv:2005.09310. doi: 10.48550/arXiv.2005.09310.
[39] L. Zhang *et al.*, "End-to-End Automatic Pronunciation Error Detection Based on Improved Hybrid CTC/Attention Architecture," *Sensors*, vol. 20, no. 7, p. 1809, Mar. 2020, doi: 10.3390/s20071809.
[40] O. Mamyrbayev, K. Alimhan, D. Oralbekova, A. Bekarystankyzy, and B. Zhumazhanov, "Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level," *EEJET*, vol. 1, no. 9(115), pp. 84–92, Feb. 2022, doi: 10.15587/1729-4061.2022.252801.
[41] C. Wang, J. Pino, and J. Gu, "Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation," Oct. 09, 2020, *arXiv*: arXiv:2006.05474. doi: 10.48550/arXiv.2006.05474.
[42] S. Guillaume *et al.*, "Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family)," in *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 170–178. doi: 10.18653/v1/2022.computel-1.21.
[43] Y. Tang, J. Pino, C. Wang, X. Ma, and D. Genzel, "A General Multi-Task Learning Framework to Leverage Text Data for Speech to Text Tasks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6209–6213. doi: 10.1109/ICASSP39728.2021.9415058.
[44] Z. Song, "English speech recognition based on deep learning with multiple features," *Computing*, vol. 102, no. 3, pp. 663–682, Mar. 2020, doi: 10.1007/s00607-019-00753-0.
[45] T. Alam, A. Khan, and F. Alam, "Punctuation Restoration using Transformer Models for High-and Low-Resource Languages," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, Online: Association for Computational Linguistics, 2020, pp. 132–142. doi: 10.18653/v1/2020.wnut-1.18.

[46] S. Wang, "Recognition of English speech – using a deep learning algorithm," *Journal of Intelligent Systems*, vol. 32, no. 1, p. 20220236, Feb. 2023, doi: 10.1515/jisys-2022-0236.
[47] J. Wang, "Speech Recognition of Oral English Teaching Based on Deep Belief Network," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 10, p. 100, Jun. 2020, doi: 10.3991/ijet.v15i10.14041.