



Review

Code-Switching in Automatic Speech Recognition: The Issues and Future Directions

Mumtaz Begum Mustafa ^{1,*} , Mansoor Ali Yusoof ², Hasan Kahtan Khalaf ³,
Ahmad Abdel Rahman Mahmoud Abushariah ⁴, Miss Laiha Mat Kiah ^{1, 4}, Hua Nong Ting ⁴
and Saravanan Muthaiyah ⁵ 

¹ Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

² Faculty of Business Finance and Information Technology, MAHSA University, Jenjarom 42610, Malaysia

³ Cardiff School of Technologies, Cardiff Metropolitan University, Llandaff Campus, Western Avenue, Cardiff CF5 2YB, UK

⁴ Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur 50603, Malaysia

⁵ Faculty of Management, BR1018, Multimedia University, Persiaran Multimedia, Cyberjaya 63100, Malaysia

* Correspondence: mumtaz@um.edu.my

Abstract: Code-switching (CS) in spoken language is where the speech has two or more languages within an utterance. It is an unsolved issue in automatic speech recognition (ASR) research as ASR needs to recognise speech in bilingual and multilingual settings, where the accuracy of ASR systems declines with CS due to pronunciation variation. There are very few reviews carried out on CS, with none conducted on bilingual and multilingual CS ASR systems. This study investigates the importance of CS in bilingual and multilingual speech recognition systems. To meet the objective of this study, two research questions were formulated, which cover both the current issues and the direction of the research. Our review focuses on databases, acoustic and language modelling, and evaluation metrics. Using selected keywords, this research has identified 274 papers and selected 42 experimental papers for review, of which 24 (representing 57%) have discussed CS, while the rest look at multilingual ASR research. The selected papers cover many well-resourced and under-resourced languages, and novel techniques to manage CS in ASR systems, which are mapping, combining and merging the phone sets of the languages experimented with in the research. Our review also examines the performance of those methods. This review found a significant variation in the performance of CS in terms of word error rates, indicating an inconsistency in the ability of ASRs to handle CS. In the conclusion, we suggest several future directions that address the issues identified in this review.

Keywords: code-switching; automatic speech recognition system; multilingual speech recognition; bilingual speech recognition; language and acoustic models; evaluation metrics



Citation: Mustafa, M.B.; Yusoof, M.A.; Khalaf, H.K.; Rahman Mahmoud Abushariah, A.A.; Kiah, M.L.M.; Ting, H.N.; Muthaiyah, S. Code-Switching in Automatic Speech Recognition: The Issues and Future Directions. *Appl. Sci.* **2022**, *12*, 9541. <https://doi.org/10.3390/app12199541>

Academic Editors: Ying Shen, Cunhang Fan and Ya Li

Received: 20 August 2022

Accepted: 20 September 2022

Published: 23 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Automatic speech recognition (ASR) is one of speech technology's most common research areas. ASR has many potentials in human-machine interaction that allow us to communicate more effectively with our devices, such as computers, robots and assistive tools for the disabled [1]. Monolingual ASR systems such as Alexa and Siri have shown the usefulness of ASR systems in our daily lives. The next apparent evolution in ASR development is in the ability to handle more than one language, as most of us can be fluent in several languages. The ability of an ASR system to handle more than one language makes our interactions with the machine "more natural".

Bilingual and multilingual speech recognition systems have many benefits, as they can be useful to recognise different languages around the globe; there are currently estimated to be around 6000 languages. They are also suitable for users who speak more than one

language. The bilingual and multilingual speech recognition systems must identify each unique language uttered by users, where speakers may be uttering more than one language in a single speech.

The notion of code-switching (CS) is common in bilingual and multilingual speech research, where the utterance includes two or more languages spoken within it [2–5]. CS creates an issue in spoken-language technologies, such as the ASR systems that need to handle input from many types of languages [6].

This study reviews the current works on CS in bilingual and multilingual ASR systems. The review focuses on the issues and solutions proposed in the selected papers in terms of the research focus, the databases, acoustic and language modelling, and evaluation metrics. The existing reviews on ASR cover mainly single language ASR systems, covering selected themes, including databases [7], feature extraction and classification [8–10] and performance [7,11], among others. At the time of writing, no works have reviewed bilingual and multilingual ASR systems, particularly on CS. This review is essential in this domain because there is an increasing need for an ASR system to recognise several languages.

The rest of the paper is organised as follows: Section 2 looks at the existing review of ASR systems and CS in bilingual and multilingual speech recognition systems. Section 3 looks at the methodology used in this research to review bilingual and multilingual speech recognition system development. The findings obtained for each research question are presented in Section 4. Section 5 discusses the significant findings. Section 6 provides the conclusion and Section 7 looks at future directions.

2. Research Background

2.1. Existing Reviews of Automatic Speech Recognition Systems

These papers were identified based on the titles, including the terms “review” or “systematic review”. The existing reviews focus on monolingual ASR systems, emphasising the machine learning approach in developing the monolingual model [12,13], noise-robust techniques for ASR systems [14], development of the end-to-end model for monolingual ASR systems [15], the error detection and corrections in monolingual ASR [11], ASR system development for various languages [16,17] and the architecture, methodology, process, databases, tools and applications of monolingual ASR [18–24]. There has been increasing interest in CS in ASR, which has resulted in many different areas of interest, solutions and outcomes. In ref. [25], the review focuses on the computational approaches for CS speech and natural language processing. However, no work summarises the progress in the development of bilingual and multilingual ASR in handling CS, allowing the researchers to better understand the progress and possible areas for future directions.

2.2. CS in Bilingual and Multilingual Speech Recognition Systems

This section examines CS in spoken language and its role in bilingual and multilingual speech recognition systems. Many current ASR systems recognise only one language (monolingual). However, a practical ASR system must be able to handle CS, which is not an easy task as the system needs to deal with multilingual input with unpredictable switching positions [26].

Individuals that can speak more than one language (e.g., bilingual or multilingual) CS or mix their languages when communicating with others [3,5,27–39]. CS is common in bilingual and multilingual communities (different cultures and language backgrounds) [40]. CS also occurs in minority languages influenced by the majority or majority languages influenced by lingua francas, such as English and French [4]. In CS, the base language refers to the language to which the syntax of a CS sentence belongs, and the foreign words are referred to as an embedded language [41].

Code-switching can occur between two languages that differ in terms of linguistic, pronunciation and speech features (for example Mandarin and English), or have high similarity in terms of linguistic, pronunciation and speech features (for example Frisian and English). CS can be classified into two primary categories: inter-sentential and intra-

sentential [2,26,42]. The classification is important as it shows where the CS will occur in a sentence as well as the speech unit involved during the CS [6].

When CS occurs between sentences, it is referred to as inter-sentential CS, where speakers use words, phrases or sentences from one language (the embedded language) with words or sentences in their primary or base language [2]. This type of CS is common among fluent bilingual speakers. For example, “If you are late for the job interview, işe alınmazsın/If you are late for the job interview, you will not be hired” (CS of English and Turkish languages), and “If I am late to the appointment, ignore sahaja/If I am late to the appointment, just ignore” (CS of Malay and English languages).

Intra-sentential CS is where the shift occurs in the middle of a sentence. For example, “Wǒ zhīdào you cóng Wǒ de péng yǒu/I know you from my friend” (CS of Mandarin and English), “Nó còng đang celebrate cái sinh nhật/He’s also celebrating his birthday” (CS of Vietnamese and English), “I tak nak you campur tangan dalam life I/I don’t want you to interfere in my life” (CS of Malay and English languages). For intra-sentential CS, the units, and the locations of the switches may vary widely from single-word switches to whole phrases (beyond the length of standard loanword units). As such, a vital issue to be solved is that the speech recognition system needs to consider many context-dependent phone combinations. However, the data sparseness problem for phone modelling hinders this issue from being solved.

The accuracy of ASR systems is adversely affected when dealing with CS speech. The reduction in accuracy is caused by unexpected pronunciation when languages are mixed. The embedded phonemes usually have more significant variation than the base language phonemes; either resembling the embedded language pronunciation or realised as a set of phonemic equivalents from the base language [2].

Early works on CS speech recognition employ the hybrid framework with three sub-models: a pronunciation model, an acoustic model and a language model [41,42]. The pronunciation model takes the pronunciation variations of words in a dictionary. On the other hand, the acoustic model is a statistical model of acoustic features for sub-word units (for example, phonemes). The language model enables the ASR system to reduce the search space by using the conditional probabilities of subsequent words with the observed word sequences. Currently, these sub-models are trained and optimised separately, leading to sub-optimal results.

The review indicates that publications on CS ASR systems are a rising trend, indicating great interest in such developments in CS, with researchers focusing on a specific issue and offering a suitable solution. Reviewing these papers can summarise issues, progress, and possible future directions in CS ASR. Thus, there is a need for a review paper that provides progress and the directions for future undertakings in this domain.

3. Research Aim and Approach

This study aims to review the current papers on CS in bilingual and multilingual ASR systems. The review covers the overview, issues, and solutions in existing works by concentrating on the research focus, the databases, language and acoustic modelling, and evaluation metrics. This review also discusses future directions in this domain.

The review process consists of three stages: formulation of research questions, search methodology, and search outcome and analysis.

3.1. Formulation of Research Questions

The research questions for this study are:

Research Question 1: What issues affect the recognition performance of CS ASR systems in bilingual and multilingual settings?

This question is answered by analysing the issues presented in the existing papers.

Research Question 2: What is the direction in the development of CS ASR systems in terms of focus, databases, acoustics and language modelling, and evaluation metrics?

This question is answered by reviewing the methods and solutions proposed in the selected papers.

3.2. Search Methodology

Here, we applied an integrated search strategy that includes searches in various online databases and a manual analysis of the selected papers. Our search covered popular databases such as:

- Science Direct;
- IEEE Explore Digital Library;
- Springer Link;
- Google Scholar.

In addition, we performed a manual review where we read the title and the abstract of the papers. We then selected the papers and discarded any irrelevant papers.

We have applied specific keywords as follows:

- multilingual speech recognition;
- bilingual speech recognition;
- code-switching.

We have applied the following inclusion criteria to screen our initial search:

- Search domain: Science, technology or computer science;
- Types of publication: Journals, proceedings, and transactions;
- Article type: Full text;
- Language: English.

We filtered irrelevant papers by referring to the following exclusion criteria:

- Papers that do not focus explicitly on bilingual and multilingual speech recognition.
- Papers that discuss bilingual and multilingual speech recognition as a side topic.
- Papers with no details of experiments or experimental design.
- The full text of the paper is not available (physical and electronic forms).
- Opinions, viewpoints, keynotes, discussions, editorials, tutorials, comments, prefaces, anecdotal papers and presentations in slide format without any associated papers.

We manually checked every identified paper to ensure that the title and abstract were relevant, and we excluded non-essential papers from indexed lists. On top of that, we have used the backward snowballing method to discover any unidentified papers from the primary strategy [43].

3.3. Search Outcome and Analysis

In this section, we present the outcome of the search and selection, and analysis of the selected experimental papers in terms of publication database sources and the languages investigated.

From the initial keyword search, we identified 274 papers. Manual analysis was performed by reading the titles and abstracts of the shortlisted papers to remove the irrelevant ones. This initial screening process eliminated 151 papers, and 123 papers remained. After reading the full papers, another 71 papers were rejected based on the exclusion criteria above, leaving 52 relevant papers. The backward snowballing method allowed us to add another 14 papers. The final number of papers for the review was 66.

In terms of sources, 27 (that is, 64%) of the selected experimental papers were from the IEEE database, 13 papers were from Google Scholar, and two papers were from the Science Direct database. In terms of the languages investigated, the selected experimental papers cover many languages, both well-resourced and under-resourced. Some of the common languages investigated in multilingual and bilingual speech recognition and CS are English (25 papers), Mandarin/Chinese (10 papers), and African languages (8 papers), among others. Figure 1 shows the languages investigated by the selected papers. Of the 42 experimental papers, half of them focused on bilingual ASR, while the other half looked at multilingual ASR. More than half of the selected papers focused on CS, with English

topping the list for CS, followed by the Chinese language. This finding is not surprising, as English is an international language. Several works focused on lesser-known languages such as isiZulu, Setswana, Frisian, and Malay, to name a few.

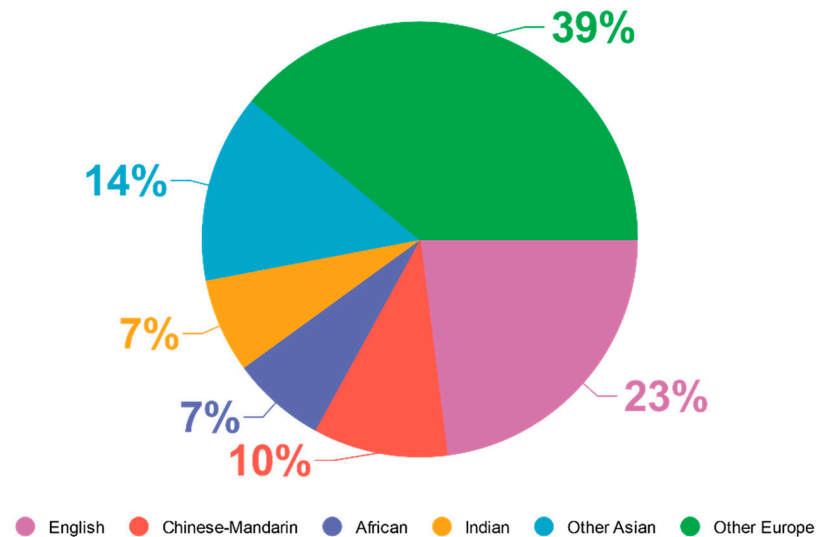


Figure 1. The languages investigated by the selected papers.

4. Results of Review

4.1. Research Question 1: What Issues Affect the Recognition Performance of CS ASR Systems in Bilingual and Multilingual Settings?

Bilingual and multilingual speech recognition and CS differ from monolingual speech recognition as they have issues and challenges. Bilingual and multilingual without CS is where a speech recognition system is not expected to recognise two or more languages in a sentence. This kind of system usually uses the existing databases for well-resourced languages [27,28,35,37,38,44–46] and is self-developed for under-resourced languages [33,34,47].

- Database Sparsity

For bilingual and multilingual speech recognition with CS, one of the significant issues is data sparsity in terms of the availability of the CS corpus and scarcity of CS occurrences in the utterances [5]. While there are established databases for monolingual speech recognition that can be used in bilingual and multilingual speech recognition, the availability of the CS corpus is limited due to the various combinations of languages in CS. A speech database covering speech variations is essential for bilingual and multilingual speech recognition [3]. Much of the research in CS develops its databases [5,29,39,41,48–50]. However, the continuing interest in CS has resulted in the development of standard databases in recent research on CS [2,4,6,30,40,47,51–55].

The existing research is focused mainly on single-language pairs of CS [26], which means bilingual speech recognition. It is because CS research aims to allow speech recognition to recognise the change in the language within a sentence uttered by multilingual users [39]. According to Nakayama et al. [26], it is prevalent that users usually switch between two languages in a sentence. However, this does not prevent researchers from investigating alternatives to modelling the acoustics for the CS of multiple language pairs, such as the universal phoneme posterior probabilities for mixed-language speech recognition, or combined language identification and speech recognition [26].

- Recognising CS

While it is challenging for CS speech recognition to recognise large output units from base and embedded languages, it is easy to recognise the specific language from a particular speech segment. However, researchers may ignore the underlying task imbalance issues when sampling tasks from CS languages, which can lead to unsatisfactory performance

by the ASR systems. First, different languages have different training data scales, so the basic task quantity for each language domain will be grossly different, leading to the task–quantity imbalance [38,53]. On top of that, as different languages have different phonological systems, the tasks taken from these languages have different recognition difficulties, causing the task–difficulty imbalance [38].

One primary concern in bilingual and multilingual speech recognition is balancing language coverage and model pollution. Using multiple source languages’ acoustic data allows for excellent context coverage. However, the differences between the base and embedded languages cause impurity in the training data, reducing the accuracy of the base language’s acoustic model. Moreover, different languages may produce mixed acoustic dynamics and context mismatch, adversely affecting the context-dependent models trained using different speech data from several languages [35].

Unlike monolingual speech recognition, CS speech is highly unpredictable and difficult to model. Despite recent advances in ASR using various machine learning algorithms, the success of CS speech was dampened due to the challenges of not having robust acoustic and language modelling that can process bilingual and multilingual speech that includes CS utterances [29,40,56].

4.2. Research Question 2: What Is the Direction in the Development of CS ASR in Terms of Focus, Databases, Acoustics and Language Modelling, and Evaluation Metrics?

Table 1 shows the direction of the selected papers in terms of focus, databases, acoustics and language modelling, and evaluation metrics.

Table 1. Direction of the papers in terms of focus, databases, acoustics and language modelling, and evaluation metrics.

Article No.	Focus		Database	Acoustics & Language Modelling				Evaluation Metrics								
	Bilingual	Multilingual		Well-Resourced	Under-Resourced	CS	Existing	Self-Develop	Mapping	Combining	Merging	Word Error Rate	Character Error Rate	CS Ratio	Mixed Error Rate/Confusion Matrix	Phoneme Error Rate
[2]	✓				✓	✓									✓	
[3]	✓						✓		✓		✓					
[4]	✓			✓	✓	✓				✓	✓					
[5]	✓	✓		✓	✓	✓				✓	✓					
[6]	✓			✓	✓		✓		✓				✓			
[26]	✓			✓	✓		✓				✓					
[27]		✓		✓		✓						✓				
[28]		✓		✓		✓			✓		✓					
[29]	✓			✓	✓	✓			✓			✓				
[30]		✓		✓	✓		✓		✓		✓					
[31]		✓		✓		✓			✓							✓
[32]	✓			✓	✓	✓								✓		
[33]		✓		✓			✓				✓					
[34]		✓		✓			✓				✓					

Table 1. Cont.

Article No.	Focus				Database		Acoustics & Language Modelling				Evaluation Metrics				
	Bilingual	Multilingual	Well-Resourced	Under-Resourced	CS	Existing	Self-Develop	Mapping	Combining	Merging	Word Error Rate	Character Error Rate	CS Ratio	Mixed Error Rate/Confusion Matrix	Phoneme Error Rate
[35]	✓		✓			✓			✓		✓				
[36]		✓		✓			✓		✓		✓				
[37]		✓	✓	✓		✓			✓		✓				
[38]		✓	✓	✓		✓				✓	✓				
[39]	✓		✓		✓	✓				✓	✓				
[40]	✓		✓		✓	✓			✓				✓	✓	
[41]	✓		✓		✓	✓			✓		✓				
[42]	✓		✓								✓				
[44]		✓	✓			✓				✓	✓				
[45]		✓	✓			✓			✓		✓				
[46]		✓	✓			✓			✓		✓				
[47]		✓	✓	✓	✓		✓		✓		✓				
[48]	✓		✓		✓		✓	✓			✓				
[49]	✓		✓		✓	✓					✓				
[50]	✓		✓		✓		✓							✓	
[51]	✓		✓		✓		✓		✓		✓				
[52]		✓	✓		✓		✓		✓		✓		✓		
[53]	✓		✓		✓		✓		✓			✓			
[54]	✓		✓				✓	✓			✓				
[55]	✓		✓		✓	✓				✓					
[56]		✓	✓		✓	✓		✓							✓
[57]		✓	✓			✓			✓		✓				
[58]	✓		✓		✓	✓			✓		✓	✓		✓	
[59]		✓	✓		✓		✓		✓		✓				
[60]		✓	✓			✓				✓	✓				
[61]		✓		✓		✓					✓				
[62]		✓	✓			✓			✓		✓				
[63]	✓		✓		✓	✓			✓					✓	
Σ	21	21	34	13	23	26	14	2	22	7	29	4	3	6	2

4.2.1. Databases

Speech recognition research has spanned several decades, and most research has focused only on a handful of languages (usually referred to as well-resourced languages) [44,45]. Many speech databases are available for these languages. The availability of databases

is not challenging in bilingual and multilingual ASR systems for well-resourced languages [28,44,45]. Bilingual and multilingual speech recognition faces challenges in the availability of databases for under-resourced languages, with many having to first develop their databases [5,34,36,39,64,65]. However, the continuing interest in CS has resulted in the development of standard databases in recent research on CS [2,4,6,30,40,47,51–55].

The existing monolingual standard databases have speech utterances in a single language. On the other hand, the CS database is a compilation of speech that contains words from two languages in the same utterance. Yilmaz et al. [39] recently compiled a CS speech corpus containing 14.3 h of language-balanced speech from soap opera broadcasts, and a Dutch broadcast database, which contained 17.5 and 89.5 h of Dutch data, respectively, while the English Broadcast News Database (HUB4) is the primary source of English broadcast data [4]. Separately, Yilmaz et al. [47] developed a bilingual lexicon containing more than 100,000 Frisian and Dutch words and about 160,000 lexicon entries due to the words with multiple phonetic transcriptions.

Of the 22 papers that investigated CS, 10 of the papers (45%) developed their CS databases, while the balance (55%) made use of the existing speech databases such as FAME (Frisian Audio Mining Enterprise) [4,5,47] and SEAME (South East Asia Mandarin–English) corpus [32,40,55,58]. Table 2 mentions details of the CS databases used in the selected research. Most CS databases cover two languages, though some databases have CS for more than two languages.

Table 2. The details of the CS databases used in the existing research.

Article No.	Database				
	Existing Database	Self-Developed	Size (HOURS)	Number of Speakers	Number of Languages
[2]	✓		NP	NP	2
[4]	✓		11.8	NP	3
[5]	✓		11.8	NP	2
[6]		✓	NP	NP	2
[26]		✓	NP	NP	3
[29]	✓		200	NP	2
[30]		✓	NP	NP	10
[32]	✓		62.8	157	2
[39]		✓	14.3	NP	6
[40]	✓		62.8	157	2
[41]	✓		25	101	2
[47]	✓		11.8	NP	2
[48]		✓	2.8	11	2
[49]	✓		20	10	2
[50]		✓	NP	NP	2
[55]	✓		62.8	157	2
[51]		✓	300	NP	2
[52]		✓	622.3	NP	10
[53]		✓	1000	NP	2
[56]	✓		6.5	143	2
[58]	✓		62.8	157	2
[59]		✓	14.3	NP	6
[63]	✓		250	NP	2

NP: not provided in the article.

Most of the databases for CS have recorded utterances of more than 10 h and involve many speakers uttering two languages. Some researchers developed the CS database artificially using the text-to-speech (TTS) system [6,26,30,50]. It can be concluded that the speech database development for CS is on par with the monolingual ASR system. Many researchers also use the existing TTS system to generate CS speech. CS involves the under-resourced languages lacking adequate databases [5,39,41]. Many of the under-resourced languages have no existing TTS system that can support the development of CS speech utterances. However, the review shows that under-resourced languages such as Sepedi have CS databases for multilingual speech recognition research [41].

4.2.2. Acoustic and Language Modelling

Most papers that investigate CS in speech are for bilingual ASR [5,6,29,32,40,41,47–51,53,55,56,58], while the rest look at multilingual ASR [4,26,30,39,52,59]. The acoustic modelling for speech recognition is a critical process in any ASR system development. Over the years, many techniques have been employed to develop the acoustic model for monolingual, bilingual, and multilingual speech recognition systems. The development of an acoustic model for the bilingual and multilingual speech recognition systems that contain CS required the model to switch language during the recognition of speech, which can happen within a sentence. The CS phone sets in bilingual and multilingual speech recognition can be made in three distinctive ways; (1) by combining the phone sets of the two languages, (2) by mapping the phone sets of the two languages, and (3) by merging similar phone sets of the two languages [2,42].

- Combining

For both the bilingual and multilingual ASR systems, researchers prefer to combine the acoustic models of each monolingual model [4,6,26,29,39–41,49,51,53,56,58,59,62]. Combining the monolingual models for CS is common for well-resourced languages such as English, Mandarin, and German, among others. Adel et al. [40] proposed the recurrent neural network language modelling toolkit for CS for bilingual ASR. Textual features such as words and part-of-speech (POS) tags were added to the input layer, while a set of all possible languages was added to the output layer. The probability for the succeeding language to be computed is based on the current word, the current features, and the history of words and features. According to Adel et al. [40], recurrent neural network language models improve the perplexity and error rates compared to the traditional n-gram approaches, as the former can handle more extended contexts. On top of that, linguistic analyses of the CS help better understand the task and challenges and thus allow researchers to develop an appropriate language model [40].

Yılmaz et al. [39] developed acoustic and language models for five languages with CS for multilingual ASR. Unlike past works that only consider bilingual CS, the developed ASR system can hypothesise CS word sequences for five languages. They use the transcriptions of the CS speech data for training the language model [39]. Alternatively, Yılmaz et al. [4] trained several multilingual deep neural network (DNN) models in Frisian, English, and Dutch for developing the CS acoustic model for low-resourced languages.

- Merging

Merging of acoustic models was applied in CS involving low-resourced languages [4,5,52]. The merging of similar phones of the two languages for CS reduces the size and complexity of the CS. For example, ref. [52] used the union character set for each language and eliminated the duplication of output symbols in multiple languages, allowing them to reduce the computational cost during model development. When recognising CS utterances, the model can switch the language of the output sequence. The bi-directional encoder network computes the hidden representations as input acoustic features, allowing the model to predict the language identification for variable-length segments [52].

In ref. [26], a spectrogram extracted by the short-time Fourier transform (STFT) was used together with the Librosa library for generating the speech features. They first

applied wave-normalisation per utterance, followed by pre-emphasis, and extracted the spectrogram with an STFT. The final set included 40 dims log mel-spectrogram features, and 1025 dims log magnitude spectrograms. On the other hand, ref. [5] applied the recurrent neural network trained with cross-lingual embedding data to maximise the use of the available textual resources.

- Mapping

Phone mapping was applied in [48], where they mapped Indonesian with Arabic phone sets. The proposed method has the advantage that the existing acoustic model with only Indonesian phones is used to build the speech recognition system. Lin et al. [57] used the universal phone set (UPS), a machine-readable phone set based on the IPA, to represent the language's universal speech units. Generally, there is a one-to-one mapping between UPS and IPA symbols, while UPS is a superset of IPA in a few other cases.

Similarly, ref. [31] proposed a method that does not require the training of language-specific acoustic models. The method created a new phoneme set to obtain multilingual acoustic modelling by sharing part of language-specific phonemes with other languages. However, sharing of phonemes increases the amount of training data. The acoustic model with the shared phoneme set can perform speech recognition for a minor (low-resource language) utterance.

- Other techniques

Recently, end-to-end (E2E) systems have been preferred by researchers for their simplicity and success in multilingual [5,29,41,46,51–53,60] settings. An E2E system directly maps an input sequence of acoustic features to an output sequence of characters, phonemes, or words. Two common variants of the E2E framework are (i) the connectionist temporal classification (CTC) [31,47], and (ii) the sequence-to-sequence modelling with an attention mechanism [41]. In E2E, the model is trained with characters as the output targets and does not include any explicit pronunciation model or language model, which means that the E2E does not need phonetically labelled training data during development. On top of that, the E2E system predicts phones or characters directly from acoustic information without any form of manually prepared alignment, making it a suitable method for multiple languages speech recognition and CS speech recognition [41].

Some common architectures for E2E are connectionist temporal classification (CTC), attention-based encoder–decoder networks, and recurrent neural network transducers [5,51,60,66]. The performance of E2E systems depends on the training data's size (it needs massive data), which can be a problem for CS for poorly resourced languages.

Seki et al. [52] introduced a hybrid attention/CTC model that used language identification explicitly and joint language identification to predict the CS between languages. Similarly, ref. [53] adopted the attention-based E2E model for the Mandarin–English CS, with three improvements, which are: (1) multi-task learning for language identity information, (2) word pieces as English modelling units to reduce the modelling unit gap between Mandarin and English, and (3) transfer learning to further improve the performance by using the larger size of Mandarin and English monolingual data. In ref. [55], the E2E and CTC were used for CS Korean–English languages.

Yue et al. [5] proposed the unsupervised two-step approach to language modelling. Unlike the traditional hybrid HMM–DNN system, an E2E CTC acoustic model is not trained using frame-level labels concerning the Cross-Entropy (CE) criterion. The model automatically learns the alignments between speech frames and label sequences using the CTC objective. This approach predicts the conditional probability of the label sequence by adding the joint probabilities of the corresponding set of CTC symbol sequences [5].

Emond et al. [30] propose the transliteration optimised word error rate to normalise the data for three types of language models, (1) a conventional n-gram language model, (2) a maximum entropy-based language model, and (3) a long short-term memory (LSTM) language model, in the CTC environment [30].

4.2.3. Evaluation Metrics

In much of the research on bilingual and multilingualism, the most common form of evaluation metrics is the word error rate (WER). This evaluation metric is common for bilingual and multilingual research for languages that use the Roman alphabet. The WER is a standard metric that allows researchers to compare their current work with previous works and provides the degree of improvement in their system's ability to recognise speech with CS. It also allows the comparison of performance among the different languages. For instance, English [28,45] performs better than other languages, and bilingual and multilingual models have a lower WER than the monolingual model [5,30,32,34,36,39,44,46,56,57,61].

However, according to Emond et al. [30], conventional WER is insufficient to measure CS languages' performance due to ambiguities in transcription, misspellings, and the use of words from two different writing systems. Such errors cause the WER of an ASR system to be higher than it should be and complicate its evaluation. On top of that, these errors make it difficult to determine the modelling errors resulting from CS language and acoustic models [30].

While WER is the primary measurement metric in ASR, some research uses character error rates (CER) [26,27,52,53], especially the research involving English–Chinese CS, as the writing symbols differ between the two languages. Mixed error rate (MER) was applied in ref. [40], which applies WER to English and CER to Mandarin.

The recognition performance of the CS bilingual and multilingual ASR systems uses the WER (30 papers), CER [27,29,30,58,59], CS ratio [6,40,52], MER/confusion matrix [2,31,32,40,50] and phoneme error rate (PER) [31,56].

We have analysed the performance of CS ASR systems in terms of their phone set arrangements as combining, mapping, or merging. Figure 2 depicts the overall results of the evaluation of ASR performance in recognising speech with CS conducted in the selected papers.

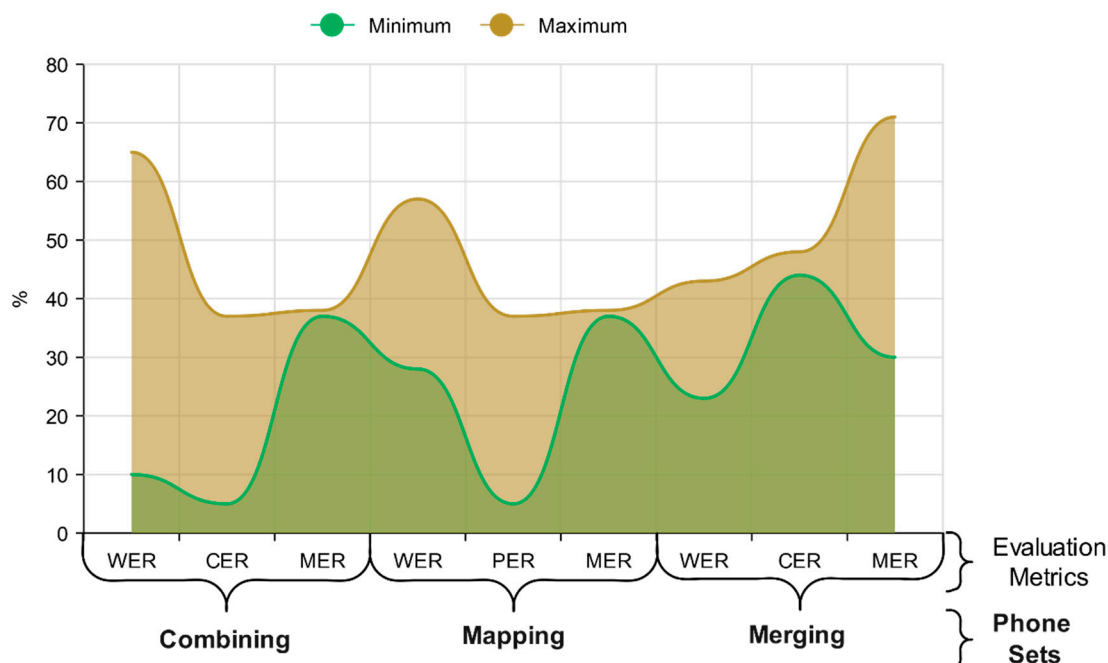


Figure 2. The overall results of the evaluation of ASR performance in recognizing speech with CS conducted in the selected papers.

As shown in Figure 2, the research that applies the combining of the phone set for CS [4,6,26,29,32,39,47,49,51,52,59] used WER, CER, and MER for measuring the performance of the ASR system. The existing work reported an error of 10% to 65% for WER, 5% to 37% for CER, and 37% to 38% for MER. The research that applied the mapping

of phone sets [5,40,48,56] has reported an error of 28% to 57% for WER, 5% to 37% for PER, and 37% to 38% for MER. Meanwhile the research on merging the phone set for CS [4,6,26,29,32,39,41,47,51,52,59], has reported an error of 23 to 43% for WER, 44% to 48% for CER, and 30% to 71% for MER.

5. Discussion

Bilingual and multilingual speech recognition with CS differs from monolingual speech recognition as it has its issues and challenges. Bilingual and multilingual recognition without CS is where the speech recognition system is not expected to switch from one language to another during the recognition process. These bilingual and multilingual systems usually use the existing databases for well-resourced languages [27,28,35,37,38,44–46,57] and are self-developed for under-resourced languages [31,32,40].

For bilingual and multilingual speech recognition with CS, 24 out of 42 (representing 57%) of experimental papers consider CS in bilingual and multilingual speech recognition development. This finding shows that CS is integral in developing bilingual and multilingual speech recognition systems. The existing research focuses mainly on single-language pairs of CS, which means bilingual speech recognition. CS research aims to allow speech recognition to recognise a change in the language within a sentence uttered by multilingual users. However, this does not prevent researchers from investigating alternatives to modelling the acoustics for the CS of multiple language pairs.

The issue of data sparsity was not new, as researchers faced the same issues in the early days of ASR research. However, today, monolingual ASR research has impressive resources in terms of databases, speakers, and recording size, enabling the development of more accurate acoustic models. Despite having many monolingual databases that support the development of bilingual and multilingual speech recognition systems, the uniqueness of the languages makes it impossible to have a common database for developing bilingual and multilingual speech recognition systems in a CS environment.

For CS, a relatively new domain in ASR research, the issue of data sparsity affects the ability of the system to recognise CS. The data sparsity includes both the availability of CS databases and the scarcity of CS occurrences in the speech corpus due to the various combinations of languages in CS. Researchers need to develop their databases in these early days, but our review shows that many now prefer to use the existing available databases instead of developing their own ones (refer to Table 2). It may indicate that developing CS databases is resource-intensive and time-consuming, where researchers prefer to use the existing CS databases for their research.

We found that the variability in the occurrence of CS is one of the reasons why CS databases are limited in terms of coverage of speech. Most CS is for bilingual ASR systems, which can have a finite mix of words from the two languages. The databases developed by researchers may not include all possible mixes, leading to sub-optimal results. The databases are also limited in terms of size and variation in speech, resulting in an imbalance of resources and tasks for developing an effective recognition model. The uneven proportion of languages in the CS database can lead to task–quantity imbalance, particularly for under-resourced languages.

Another issue in CS is the need for manual annotation of the CS database, causing the development of the CS database to be expensive and time-consuming. A possible solution to manual annotation is to use the E2E method that directly converts input speech feature sequences to output label sequences without any explicit and intermediate representation of phonetic/linguistic constructs such as phonemes or words. Some researchers developed the CS database using recorded speech from sources such as broadcasts and podcasts. This form of database offers more natural CS as compared to those recorded in the studio. The main issue will be the resources needed to annotate these speeches. However, a method such as E2E allows the development of acoustic models without the need for manual annotation. Enriching the CS database from existing sources and finding an effective way to annotate speech can be one of the future directions in CS research.

A critical process in any speech recognition system development is the acoustic modelling of speech. Over the years, many techniques have been employed to develop the acoustic model for monolingual, bilingual, and multilingual speech recognition systems. The development of the acoustic model for the bilingual and multilingual speech recognition systems that contain CS requires the model to switch languages during the recognition of speech, which can happen at any point within a sentence. The ability of the ASR system to identify the CS occurrence determines the recognition performance of the system. This can be challenging due to the variation in ways a speaker can CS during a conversation. As such, the ASR system requires sufficient coverage of CS speech in real life. Using naturally recorded speech such as broadcasts and podcasts is one possible solution. In addition, having the ASR system predict the occurrence of CS in speech from the linguistic point of view will help the ASR system to improve its recognition performance.

The CS for languages that differ in terms of linguistic, pronunciation and speech features can be better predicted for CS occurrence than for languages that have high similarity in terms of linguistic, pronunciation and speech features. The use of semantic analysis may help ASR systems to improve their CS recognition performance for languages that share similar linguistic, pronunciation and speech features.

Most papers use supervised and semi-supervised training to develop acoustic and language models. Deep learning algorithms such as the neural network and recurrent neural network language modelling are common for CS. Existing techniques such as the neural network reduce the sample size and the acoustic model's complexity for the monolingual ASR system. However, this can be a problem in CS recognition due to the unification of similar vocabulary for both languages, resulting in lower recognition accuracy. The development of CS ASR systems needs to balance language coverage and model complexity.

The CS phone sets in bilingual and multilingual speech recognition can be made in three distinctive ways: combining, mapping, and merging similar phone sets of the languages. Researchers prefer combining the acoustic models of each monolingual model for both the bilingual and multilingual ASR systems. Combining the monolingual models for CS is typical for well-resourced languages. Merging of acoustic models was applied in CS involving low-resourced languages [4,5,52]. The merging of similar phones of the two languages for CS reduces the size and complexity of the CS. Mapping was applied in some research where the languages share common graphemes, such as Indonesian and Arabic.

In the earlier section, we presented the overall results of the evaluation of ASR performance in recognising speech with CS, conducted in the selected papers. Unlike monolingual ASR, where the standard evaluation metric is the WER, the CS evaluation uses several different measurements, such as the PER, CER, and MER, depending on the type of languages being tested. For example, the WER is typical for languages that use the Roman alphabet in many European countries, while the CER is ubiquitous in CS for English and Mandarin, with different written formats.

In ASR research, it is difficult to make an “apples to apples” comparison between the methods due to variations in the inputs such as database size and the number of speakers, among others. The performance of ASR in recognising CS may not be entirely due to the techniques used for training the models, but also to the size and the quality of the database. It is important for researchers to differentiate the improvement to the WER that is caused by the proposed technique as well as the quality of the database used.

We also found that depending entirely on the WER as the standard measure of performance of ASR systems may not give readers the complete picture of the merits and demerits of a particular technique applied in developing an ASR system for handling CS. The researchers should consider other measures such as computational time, computational cost, line of code, human cost and time, and resource costs when evaluating the performance of bilingual and multilingual ASR systems.

Recently, E2E systems have been preferred by researchers for their simplicity and success in bilingual and multilingual settings. An E2E system predicts phones or characters

directly from acoustic information without predefined alignment. Unlike the traditional hybrid HMM–DNN system, an E2E CTC acoustic model is not trained using frame-level labels, reducing human involvement during training. However, the performance of E2E systems depends on the training data size (it needs massive data), which can be a problem for CS or poorly resourced languages. The performance of E2E, though promising for monolingual, bilingual, and multilingual ASR systems, may not be a suitable solution in CS due to the limited size of the CS database.

6. Conclusions

CS is an integral part of developing bilingual and multilingual speech recognition systems. Much of the speech recognition system development now focused on the bilingual and multilingual as monolingual speech recognition system has been a commercial success. As humans are commonly fluent in two or more languages, so should the ASR system be. However, unlike humans, who can naturally change from one language to another quickly, most of the existing ASR systems do not have similar abilities, particularly in handling unexpected changes in pronunciation when languages are mixed. CS may involve two languages that differ in terms of linguistic, pronunciation, and speech features or have a high degree of similarity between them. In both situations, the ASR systems need to differentiate the two languages from the uttered speech.

The aim of this research is to perform a review on CS in bilingual and multilingual ASR systems. This is because there is very little research that summarises the progress in the development of bilingual and multilingual ASR systems that can handle CS with acceptable performance.

To meet the objective of this study, two research questions were formulated. The first research question identifies the issues affecting the recognition performance of CS ASR systems in bilingual and multilingual settings. For this, we reviewed 66 selected papers (both review and experimental papers), where the issues in CS recognition can be grossly categorized into two, which are database sparsity and recognising CS. The database sparsity is mainly caused by the variability in the occurrence of CS, which is not fully covered by the existing CS databases. Recognising CS is difficult as CS speech is highly unpredictable and difficult to model.

The second research question looks at the direction in the development of CS ASR systems in terms of focus, databases, acoustics and language modelling, and evaluation metrics. For this, we reviewed 42 selected experimental papers (reflected in Table 1). These papers cover many well-resourced and under-resourced languages and techniques to recognise CS in ASR systems, such as mapping, combining, and merging the phone sets of the languages experimented with. Our review also examined the performance of those techniques. This review found a significant variation in the performance of CS experimental papers in terms of WER, indicating the inconsistency in the ability of the existing ASR systems to handle CS.

7. Future Directions

The future direction of CS will include more languages as globalisation continues. Our review shows that researchers prefer to use the existing CS databases in their research, which means that more CS database enrichment and unification will be needed in the future. Recorded speech from sources such as broadcasts and podcasts offer more natural CS as compared to those recorded in the studio. As such, the CS database of the future will use the existing recordings using automatic annotating methods. On top of that, methods such as E2E allow the development of an acoustic model without the need for manual annotation, and facilitate the development of CS databases from natural sources such as recorded broadcasts or podcasts. The development of a common CS database will help researchers in the improvement of the performance of ASR for bilingual and multilingual speech. Furthermore, there will be more development of multilingual CS in multi-ethnic communities worldwide, including for under-resourced languages.

The second issue identified in the review is recognizing CS, which can lead to unsatisfactory performance by the ASR systems. Different languages have different training data scales and phonological systems, making it difficult for the ASR system to recognise CS speech. One possible future direction for recognizing CS is to develop an ASR system that can predict the occurrence of CS in speech using linguistic information, as well as semantic analysis, to improve CS recognition performance for languages that share similar linguistic, pronunciation, and speech features.

While the standard evaluation metric for the ASR system is WER, researchers may need to differentiate or identify the errors that arose from the CS with the speech recognition error by the acoustic model. The performance of the ASR in recognising CS may not be entirely due to the techniques used for training the models, but also to the database used. In the future, researchers need to differentiate the improvement to the WER that is caused by the proposed technique as well as the quality of the database used. Such identification can help researchers to determine which factor(s) contribute more to the performance improvement.

Author Contributions: Conceptualization, methodology, writing—original draft, results analysis, M.B.M.; data collection, data analysis, writing—review and editing, results analysis, M.A.Y.; methodology, writing—review and editing, design and presentation, references, H.K.K.; technical content, writing—review and editing, A.A.R.M.A.; methodology, writing—review and editing, M.L.M.K.; technical content, writing—review and editing, H.N.T.; methodology, writing—review and editing, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/1/2020/ICT09/UM/02/1).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Al-Qatab, B.A.; Mustafa, M.B. Classification of Dysarthric Speech According to the Severity of Impairment: An Analysis of Acoustic Features. *IEEE Access* **2021**, *9*, 18183–18194. [\[CrossRef\]](#)
2. Modipa, T.I.; Davel, M.H. Predicting vowel substitution in code-switched speech. In Proceedings of the Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), Port Elizabeth, South Africa, 26–27 November 2015.
3. Shen, H.-P.; Wu, C.-H.; Yang, Y.-T.; Hsu, C.-S. CECOS: A Chinese-English code-switching speech database. In Proceedings of the International Conference on Speech Database and Assessments (Oriental COCODA), Hsinchu, Taiwan, 26–28 October 2011.
4. Yilmaz, E.; van den Heuvel, H.; Van Leeuwen, D. Code-switching detection using multilingual DNNs. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016.
5. Yue, X.; Lee, G.; Yilmaz, E.; Deng, F.; Li, H. End-to-end code-switching ASR for low-resourced language pairs. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019.
6. Nakayama, S.; Tjandra, A.; Sakti, S.; Nakamura, S. Speech chain for semi-supervised learning of Japanese-English code-switching ASR and TTS. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018.
7. Alharbi, S.; Alrazgan, M.; Alrashed, A.; AlNomasi, T.; Almojel, R.; Alharbi, R.; Alharbi, S.; Alturki, S.; Alshehri, F.; Almojil, M. Automatic speech recognition: Systematic literature review. *IEEE Access* **2021**, *9*, 131858–131876. [\[CrossRef\]](#)
8. Bell, P.; Fainberg, J.; Klejch, O.; Li, J.; Renals, S.; Swietojanski, P. Adaptation algorithms for neural network-based speech recognition: An overview. *IEEE Open J. Signal Process.* **2020**, *2*, 33–66. [\[CrossRef\]](#)
9. Desai, N.; Dhameliya, K.; Desai, V. Feature extraction and classification techniques for speech recognition: A review. *Int. J. Emerg. Technol. Adv. Eng.* **2013**, *3*, 367–371.
10. Sarma, M.; Sarma, K.K. Acoustic modeling of speech signal using artificial neural network: A review of techniques and current trends. In *Intelligent Applications for Heterogeneous System Modeling and Design*; IGI Global: Hershey, PA, USA, 2015; pp. 282–299.
11. Errattahi, R.; El Hannani, A.; Ouahmane, H. Automatic speech recognition errors detection and correction: A review. *Procedia Comput. Sci.* **2018**, *128*, 32–37. [\[CrossRef\]](#)

12. Padmanabhan, J.; Johnson Premkumar, M.J. Machine learning in automatic speech recognition: A survey. *IETE Tech. Rev.* **2015**, *32*, 240–251. [\[CrossRef\]](#)
13. Deng, L.; Li, X. Machine learning paradigms for speech recognition: An overview. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1060–1089. [\[CrossRef\]](#)
14. Li, J.; Deng, L.; Gong, Y.; Haeb-Umbach, R. An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 745–777. [\[CrossRef\]](#)
15. Wang, D.; Wang, X.; Lv, S. An overview of end-to-end automatic speech recognition. *Symmetry* **2019**, *11*, 1018. [\[CrossRef\]](#)
16. de Lima, T.A.; Da Costa-Abreu, M. A survey on automatic speech recognition systems for Portuguese language and its variations. *Comput. Speech Lang.* **2020**, *62*, 101055. [\[CrossRef\]](#)
17. Singh, A.; Kadyan, V.; Kumar, M.; Bassan, N. ASRoIL: A comprehensive survey for automatic speech recognition of Indian languages. *Artif. Intell. Rev.* **2020**, *53*, 3673–3704. [\[CrossRef\]](#)
18. Ghai, W.; Singh, N. Literature review on automatic speech recognition. *Int. J. Comput. Appl.* **2012**, *41*, 42–50. [\[CrossRef\]](#)
19. Aldarmaki, H.; Ullah, A.; Ram, S.; Zaki, N. Unsupervised automatic speech recognition: A review. *Speech Commun.* **2022**, *139*, 76–91. [\[CrossRef\]](#)
20. Anusuya, M.; Katti, S. Front end analysis of speech recognition: A review. *Int. J. Speech Technol.* **2011**, *14*, 99–145. [\[CrossRef\]](#)
21. Arora, S.J.; Singh, R.P. Automatic speech recognition: A review. *Int. J. Comput. Appl.* **2012**, *60*, 34–44.
22. Cutajar, M.; Gatt, E.; Grech, I.; Casha, O.; Micallef, J. Comparative study of automatic speech recognition techniques. *IET Signal Process.* **2013**, *7*, 25–46. [\[CrossRef\]](#)
23. Karpagavalli, S.; Chandra, E. A review on automatic speech recognition architecture and approaches. *Int. J. Signal Process. Image Process. Pattern Recognit.* **2016**, *9*, 393–404.
24. Young, V.; Mihailidis, A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assist. Technol.* **2010**, *22*, 99–112. [\[CrossRef\]](#)
25. Sitaram, S.; Chandu, K.R.; Rallabandi, S.K.; Black, A.W. A survey of code-switched speech and language processing. *arXiv* **2019**, arXiv:1904.00784.
26. Nakayama, S.; Tjandra, A.; Sakti, S.; Nakamura, S. Zero-shot code-switching ASR and TTS with multilingual machine speech chain. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019.
27. Chen, Y.-C.; Hsu, J.-Y.; Lee, C.-K.; Lee, H.-Y. DARTS-ASR: Differentiable architecture search for multilingual speech recognition and adaptation. *arXiv* **2020**, arXiv:2005.07029.
28. Biswas, A.; Yilmaz, E.; De Wet, F.; van der Westhuizen, E.; Niesler, T. Semi-supervised development of ASR systems for multilingual code-switched speech in under-resourced languages. *arXiv* **2020**, arXiv:2003.03135.
29. Du, C.; Li, H.; Lu, Y.; Wang, L.; Qian, Y. Data augmentation for end-to-end code-switching speech recognition. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021.
30. Emond, J.; Ramabhadran, B.; Roark, B.; Moreno, P.; Ma, M. Transliteration-based approaches to improve code-switched speech recognition performance. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018.
31. Hara, S.; Nishizaki, H. Acoustic modeling with a shared phoneme set for multilingual speech recognition without code-switching. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017.
32. Huang, Z.; Li, P.; Xu, J.; Zhang, P.; Yan, Y. Context-dependent Label Smoothing Regularization for Attention-based End-to-End Code-Switching Speech Recognition. In Proceedings of the 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), Hong Kong, China, 24–27 January 2021.
33. Imseng, D.; Bourslard, H.; Garner, P.N. Using KL-divergence and multilingual information to improve ASR for under-resourced languages. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012.
34. Kannan, A.; Datta, A.; Sainath, T.N.; Weinstein, E.; Ramabhadran, B.; Wu, Y.; Bapna, A.; Chen, Z.; Lee, S. Large-scale multilingual speech recognition with a streaming end-to-end model. *arXiv* **2019**, arXiv:1909.05330.
35. Lin, H.; Deng, L.; Yu, D.; Gong, Y.-f.; Acero, A.; Lee, C.-H. A study on multilingual acoustic modeling for large vocabulary ASR. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009.
36. Liu, C.; Zhang, Q.; Zhang, X.; Singh, K.; Saraf, Y.; Zweig, G. Multilingual graphemic hybrid ASR with massive data augmentation. *arXiv* **2019**, arXiv:1909.06522.
37. Pratap, V.; Sriram, A.; Tomasello, P.; Hannun, A.; Liptchinsky, V.; Synnaeve, G.; Collobert, R. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *arXiv* **2020**, arXiv:2007.03001.
38. Xiao, Y.; Gong, K.; Zhou, P.; Zheng, G.; Liang, X.; Lin, L. Adversarial meta sampling for multilingual low-resource speech recognition. *arXiv* **2020**, arXiv:2012.11896.
39. Yilmaz, E.; Biswas, A.; van der Westhuizen, E.; de Wet, F.; Niesler, T. Building a unified code-switching ASR system for South African languages. *arXiv* **2018**, arXiv:1807.10949.

40. Adel, H.; Vu, N.T.; Kraus, F.; Schlippe, T.; Li, H.; Schultz, T. Recurrent neural network language modeling for code switching conversational speech. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013.
41. Sreeram, G.; Sinha, R. Exploration of end-to-end framework for code-switching speech recognition task: Challenges and enhancements. *IEEE Access* **2020**, *8*, 68146–68157. [[CrossRef](#)]
42. Wu, C.-H.; Shen, H.-P.; Yang, Y.-T. Chinese-English phone set construction for code-switching ASR using acoustic and DNN-extracted articulatory features. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 858–862. [[CrossRef](#)]
43. Petersen, K.; Vakkalanka, S.; Kuzniarz, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Inf. Softw. Technol.* **2015**, *64*, 1–18. [[CrossRef](#)]
44. Tong, S.; Garner, P.N.; Bourlard, H. Multilingual training and cross-lingual adaptation on CTC-based acoustic model. *arXiv* **2017**, arXiv:1711.10025.
45. Tüske, Z.; Schlüter, R.; Ney, H. Multilingual hierarchical MRASTA features for ASR. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013.
46. Zhou, S.; Xu, S.; Xu, B. Multilingual end-to-end speech recognition with a single transformer on low-resource languages. *arXiv* **2018**, arXiv:1806.05059.
47. Yilmaz, E.; McLaren, M.; van den Heuvel, H.; van Leeuwen, D.A. Language diarization for semi-supervised bilingual acoustic model training. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017.
48. Barik, R.E.; Lestari, D.P. Text corpus and acoustic model addition for Indonesian-Arabic code-switching in automatic speech recognition system. In Proceedings of the International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Yogyakarta, Indonesia, 20–21 September 2019.
49. Masekwameng, M.S.; Mokgonyane, T.B.; Modipa, T.I.; Manamela, M.J.; Mogale, M.M. Effects of Language Modelling for Sepedi-English Code-Switched Speech in Automatic Speech Recognition System. In Proceedings of the International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 6–7 August 2020.
50. Shah, S.; Sitaram, S. Using monolingual speech recognition for spoken term detection in code-switched hindi-english speech. In Proceedings of the 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 8–11 November 2019.
51. Li, K.; Li, J.; Ye, G.; Zhao, R.; Gong, Y. Towards code-switching ASR for end-to-end CTC models. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
52. Seki, H.; Watanabe, S.; Hori, T.; Le Roux, J.; Hershey, J.R. An end-to-end language-tracking speech recognizer for mixed-language speech. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
53. Shan, C.; Weng, C.; Wang, G.; Su, D.; Luo, M.; Yu, D.; Xie, L. Investigating end-to-end speech recognition for mandarin-english code-switching. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
54. Vu, N.T.; Lyu, D.-C.; Weiner, J.; Telaar, D.; Schlippe, T.; Blaicher, F.; Chng, E.-S.; Schultz, T.; Li, H. A first speech recognition system for Mandarin-English code-switch conversational speech. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012.
55. Lee, D.; Kim, D.; Yun, S.; Kim, S. Phonetic Variation Modeling and a Language Model Adaptation for Korean English Code-Switching Speech Recognition. *Appl. Sci.* **2021**, *11*, 2866. [[CrossRef](#)]
56. Mabokela, K.R. A multilingual ASR of Sepedi-English code-switched speech for automatic language identification. In Proceedings of the 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Vanderbijlpark, South Africa, 21–22 November 2019.
57. Lin, H.; Deng, L.; Droppo, J.; Yu, D.; Acero, A. Learning methods in multilingual speech recognition. In Proceedings of the Proc. NIPS, Vancouver, BC, Canada, 12–13 December 2008.
58. Song, X.; Zou, Y.; Huang, S.; Chen, S.; Liu, Y. Investigating multi-task learning for automatic speech recognition with code-switching between Mandarin and English. In Proceedings of the 2017 International Conference on Asian Language Processing (IALP), Singapore, 5–7 December 2017.
59. Biswas, A.; de Wet, F.; van der Westhuizen, E.; Yilmaz, E.; Niesler, T. Multilingual Neural Network Acoustic Modelling for ASR of Under-Resourced English-isiZulu Code-Switched Speech. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018.
60. Tong, S.; Garner, P.N.; Bourlard, H. An investigation of multilingual ASR using end-to-end LF-MMI. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
61. Toshniwal, S.; Sainath, T.N.; Weiss, R.J.; Li, B.; Moreno, P.; Weinstein, E.; Rao, K. Multilingual speech recognition with a single end-to-end model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
62. Müller, M.; Stiiker, S.; Waibel, A. Multilingual adaptation of RNN based ASR systems. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
63. Song, T.; Xu, Q.; Ge, M.; Wang, L.; Shi, H.; Lv, Y.; Lin, Y.; Dang, J. Language-specific Characteristic Assistance for Code-switching Speech Recognition. *arXiv* **2022**, arXiv:2206.14580.

-
64. Mustafa, M.B.; Mohd Don, Z.; Ainon, R.; Zainuddin, R.; Knowles, G. Developing an HMM-Based Speech Synthesis System for Malay: A Comparison of Iterative and Isolated Unit Training. *IEICE Trans. Inf. Syst.* **2014**, *97*, 1273–1282. [[CrossRef](#)]
 65. Mustafa, M.; Ainon, R. Emotional speech acoustic model for Malay: Iterative versus isolated unit training. *J. Acoust. Soc. Am.* **2013**, *134*, 3057–3066. [[CrossRef](#)]
 66. Huang, Z.; Wang, P.; Wang, J.; Miao, H.; Xu, J.; Zhang, P. Improving Transformer Based End-to-End Code-Switching Speech Recognition Using Language Identification. *Appl. Sci.* **2021**, *11*, 9106. [[CrossRef](#)]