applied sciences

MDPI

Article

Phonetic Variation Modeling and a Language Model Adaptation for Korean English Code-Switching Speech Recognition

Damheo Lee [1], Donghyun Kim [2], Seung Yun [2] and Sanghun Kim [2,*]

[1] Software Development Department, IIR TECH Inc., Daejeon 34134, Korea; leedheo@iirtech.co.kr
[2] Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon 34129, Korea; dawnkann@etri.re.kr (D.K.); syun@etri.re.kr (S.Y.)
* Correspondence: ksh@etri.re.kr; Tel.: +82-42-860-5141

Featured Application: Korean English Code-Switching.

Abstract: In this paper, we propose a new method for code-switching (CS) automatic speech recognition (ASR) in Korean. First, the phonetic variations in English pronunciation spoken by Korean speakers should be considered. Thus, we tried to find a unified pronunciation model based on phonetic knowledge and deep learning. Second, we extracted the CS sentences semantically similar to the target domain and then applied the language model (LM) adaptation to solve the biased modeling toward Korean due to the imbalanced training data. In this experiment, training data were AI Hub (1033 h) in Korean and Librispeech (960 h) in English. As a result, when compared to the baseline, the proposed method improved the error reduction rate (ERR) by up to 11.6% with phonetic variant modeling and by 17.3% when semantically similar sentences were applied to the LM adaptation. If we considered only English words, the word correction rate improved up to 24.2% compared to that of the baseline. The proposed method seems to be very effective in CS speech recognition.

Keywords: speech recognition; code-switching; language model; domain adaptation; acoustic model; shallow fusion

check for updates

Citation: 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. https://doi.org/10.3390/app11062866

Academic Editor: José A. González-López

Received: 16 February 2021
Accepted: 17 March 2021
Published: 23 March 2021

1. Introduction

Recently, automatic speech recognition (ASR) and speech translation (ST) based on end-to-end (E2E) frameworks have shown significant improvements. These systems have been widely adapted to real-life situations, such as lectures, business meetings, and human–machine conversations. Figure 1 shows an application for English–Korean ASR.

However, as the use of foreign words is more common these days, which tends to frequently cause accuracy degradation. Many researchers have studied this issue, which is called the code-switching (CS) problem in ASR. In the case of Korean, English words pronounced by Korean speakers—Korean-style English (i.e., Konglish)—have many phonetic variations from native-like English pronunciation. Therefore, reflecting proper phonetic variations with conventional methods is very complex. Moreover, mixed-language spoken data are very rare, making any model biased toward Korean, even if there are many data, so a sophisticated approach is needed.

To figure out the effect of CS, we investigated how often Korean sentences have English words. In the broadcasting news domain, 26 million (2.1%) out of 1277 million words are from English. The IT domain is more severe, as can be expected. There are 127 million (11.5%) English words out of 1011 million words in that domain. Figure 2 shows typical CS sentences in Korean.

Generally, these problems can be categorized into two types: the inter-sentential, where language transitions occur at the phrase, sentence, or discourse boundaries; and the intra-sentential, where language transitions occur without the interruption of speech in the middle of sentences [1].

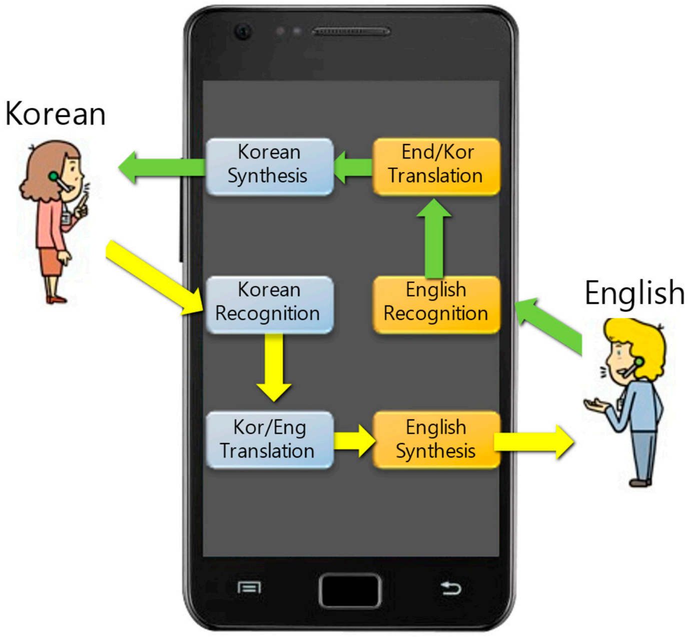
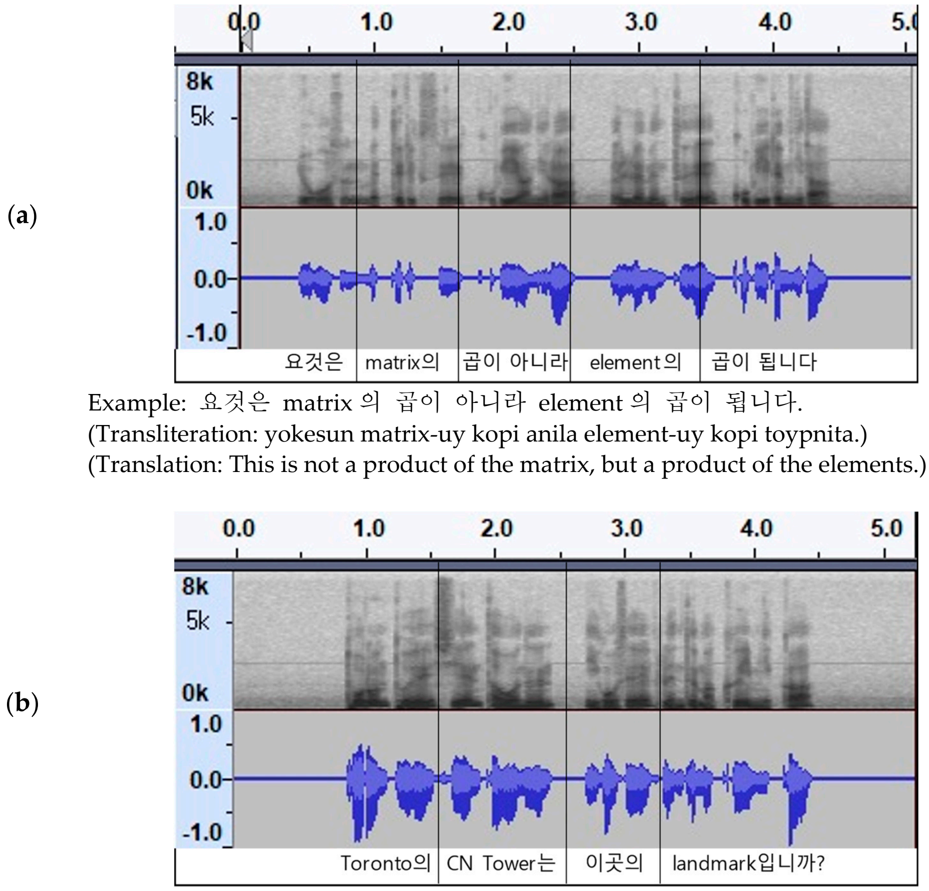


Figure 1. An application of the proposed method: English–Korean automatic speech translator.



Example: 요것은 matrix 의 곱이 아니라 element 의 곱이 됩니다.
(Transliteration: yokesun matrix-uy kopi anila element-uy kopi toypnita.)
(Translation: This is not a product of the matrix, but a product of the elements.)

Example: Toronto 의 CN Tower 는 이곳의 landmark 입니까?
(Transliteration: Toronto-uy CN tower-nun igosuy landmark-ipnikka?)
(Translation: Is the CN Tower in Toronto the landmark here?)

Figure 2. Typical code-switching (CS) sentences: (**a**) lecture domain; (**b**) travel domain.

In this paper, we focus on intra-sentential CS problems. In Section 2, we introduce the CS research results obtained in other studies, and Section 3 explains the difficulties in modeling phonetic variations in English spoken by Koreans. Section 4 handles how to extract sentences with similar meanings for the language model (LM) domain adaptation. Section 5 summarizes the experimental results for the proposed method and concludes the paper.

2. Related Work

For a long time, language-specific speech recognition with language identification tags [2] has been studied as an intuitive approach. To detect language boundaries, the bi-phonemic probability can be calculated [3], which measures the confidence score of the phoneme using a foreign language in a CS sentence. Watanabe et al. [4] and Seki et al. [5] adopted language tags in sentence units in ASR to train the model with a language tag to distinguish language-specific characteristics.

As another approach, the context-switching database (DB) was directly built by mixing several languages to solve the data imbalance [6]. Some studies approached overcoming low-resource language pairs, which relate to the data imbalance for ASR [7–9]. Similarly, a study was conducted to create a more robust model using an asymmetric corpus by language [10].

Among the well-known studies, one used transliteration with Latin characters. Vu et al. [11] proposed a knowledge-based phoneme merging or data-driven merging in Chinese English, and as a similar approach, training with code-mixed resources in Hindi English was attempted [12,13]. Recently, data augmentation was used for generating a mixed corpus with a generative adversarial network (GAN) using a monolingual corpus and a few CS sentences [14]. According to Long et al. [15], the acoustic data augmentation was accomplished in a English–Chinese CS speech recognition task. The unsupervised learning technique was also utilized with a monolingual CS corpus [16].

3. Phonetic Variant Modeling

Gathering a CS corpus for training is ineffective because its model inclines toward Korean as the data size grows. For this reason, SEAME [6], a Mandarin–English code-switching speech corpus in Southeast Asia, built a corpus using mixed sentences specially designed for Chinese English. Tjandra et al. [17] proposed the speech chain algorithm, which synthesizes speech from recognized text and then feeds it back again. Nakayama et al. [18,19] improved the performance by generating Japanese English intra-sentential sentences. In this study, this corpus was generated by substituting katakana words or phrases with English words.

To simultaneously avoid the data imbalance and low resources of CS, in this paper, we propose a hybrid method based on phonetic knowledge and deep learning, which integrates Korean and English data. For this, defining a unified alphabet was necessary to solve the intra-sentential CS problem. As a first step, linguistic or phonetic knowledge was introduced to map English phonemes to Korean phonemes based on the phonetic similarity between the two languages. Second, after applying mapping rules, end-to-end ASR was trained with reference to natural Konglish pronunciations.

3.1. *Phoneme Mapping Using Phonetic Knowledge*

Table 1 [20] shows the consonants in Korean and English that are related to the places of articulation, which describe the movements of the mouth, teeth, tongue, or vocal tract. English phonemes include the labial fricative /v/, dental fricative /ð/, and alveolar approximant /r/, which do not exist in Korean. If these phonemes are approximated to Korean phonemes /p/, /t/, and /l/, any English words can be transliterated into Hangul. However, phoneme differences (i.e., acoustic differences) still exist between Korean-style English (i.e., Konglish) and native English.

Table 1. Difference of consonants map: (**a**) Korean [1]; (**b**) American English.

(**a**)

	Labial	**Dental**	**Alveolar**	**Post-alveolar**	**Palatal**	**Velar**	**Glottal**
Nasal	/m/		/n/			/ŋ/	
Affiricate Fortis	/p/		/t/	/c/		/k/	
Aff. Fortis Tense	/p′/		/t′/	/c′/		/k′/	
Aff. Fortis Aspirated	/p^{h}/		/t^{h}/	/c^{h}/		/k^{h}/	
Fricative Fortis			/s/				/h/
Fricative Tense			/s′/				
Approximant			/l/				

(**b**)

	Labial	**Dental**	**Alveolar**	**Post-alveolar**	**Palatal**	**Velar**	**Glottal**
Nasal	/m/		/n/			/ŋ/	
Affiricate Fortis	/p/		/t/	/tʃ/		/k/	
Aff. Fortis Lenis	/b/		/d/	/dʒ/		/g/	
Fricative Fortis	/f/	/θ/	/s/	/ʃ/			/h/
Fricative Lenis	/v/	/ð/	/z/	/ʒ/			
Approximant			/l/	/r/	/j/	/w/	

[1] These consonants are different from the pronunciation-based notation (e.g., /c/, /p/, and /l/ in Korean).

Unlike other languages, Konglish seems to be more severe in its phonetic variations. We should consider the phonetic variations between English pronounced by a Korean who has difficulty speaking English and English pronounced by a Korean who speaks English at a native-like level. In the former case, English phonemes are transformed into Korean-style English phonemes [21,22] (i.e., Valentine's /vælntanz/ is changed into 발렌타인스 /ballntains/); in the latter case, the phonemes are closer to the native English phonemes. (i.e., Valentine's /vælntaInz/ is changed into 밸런타인즈 /bllntainc/). Thus, due to the abundant phonetic variance between Konglish and native English, it is difficult to make acoustic models for CS speech recognition.

Table 2 shows the phoneme relationship between Korean (KR) and English (EN) based on phonetic knowledge. Some English phonemes can be directly mapped into Korean phonemes, but others should be approximately mapped to similar phonemes in Korean where possible. In this study, we defined English phonemes based on CMUdict [23] and applied the Korean English pronunciation conversion rule [24]. For example, to map the English phoneme /k/ onto the Korean phoneme /k/ or /G/, the Korean phoneme /k/ is used when it is placed at the beginning of a word, and /G/ is used when it placed amid a word. Several English monophthong or diphthong vowels should be forcedly approximated as Korean phoneme /v/. The following examples show the results based on the rules:

- Examples: Access rights, scratch language, and Taylor Swift.
- After phoneme mapping: 액세스 라이츠, 스크래치 랭귀지, 앤드 테일러 스위프트. (Transliteration: /aykseysu laichu, sukhulaychi layngkwici, ayntu theyille suwiphuthu/.)
- After applying Konglish rules: 액셋 롸잇츠, 스크래치 랭귀쥐, 앤 테일르 스웝트. (Transliteration: /aykset lwaitchu, sukhulaychi layngkwicyu, ayn thayllu suwipthu/.)

Indeed, the Korean language is composed of syllable structures as pronunciation units, so vowel insertion occurs between consonants. Consecutive consonants in an English word should be used to form a syllable structure. For this, a Konglish dictionary with regular conversion rules was made. The results show that the rules work well. To produce a Konglish dictionary, Phonetisaurus [25] was adopted—the English grapheme-to-phoneme (G2P) toolkit. For instance, with the conversion rules in Table 2, "school" becomes /skul/

in English G2P, and converts to Konglish phoneme sequence /sUkul/, again via the conversion rules. Then, /sUkul/ is simply transformed to "스쿨(*/sukhul/*)" through Hangul conversion.

Table 2. Phoneme conversion between English and Korean.

EN [1]	**KR** [2]	**Han** [3]	**EN** [1]	**KR** [2]	**Han** [3]	**EN** [1]	**KR** [2]	**Han** [3]
b/v	b	ㅂ	t / th	t/D	ㅌ / ㄸ	w ae	wE	ㅙ
ch	c/Z	ㅊ / ㅉ	ng	N	ㅇ	w ah	wv	ㅝ
d/dh	d	ㄷ	aa	a	ㅏ	w eh	we	ㅞ
g	g	ㄱ	ae	E	ㅐ	w iy	wi	ㅟ
hh	h	ㅎ	ah / eh uh / iy ah	v	ㅓ	y aa	ja	ㅑ
jh/z/zh	z	ㅈ	eh	e	ㅔ	y ae	jE	ㅒ
k	k/G	ㅋ / ㄲ	ih / uh ih	Wi	ㅢ	y ah	jv	ㅕ
m	m	ㅁ	iy	i	ㅣ	y eh	je	ㅖ
n	n	ㄴ	ow / ao	o	ㅗ	y ow/y ao	jo	ㅛ
p/f	p/B	ㅍ / ㅃ	uh / ih uh	U	ㅡ	y uw	ju	ㅠ
r/l	r	ㄹ	uw	u	ㅜ			
s/sh	s/S	ㅅ / ㅆ	w aa	wa	ㅘ			

[1] English phoneme based on CMUdict. [2] Transliterated Korean phoneme with Hangul. [3] Hangul character equivalent to each Korean phoneme.

3.2. Considering Phonetic Variations Using End-to-End ASR

Until now, we have dealt with the phonetic modeling of English pronounced by a Korean who has difficulty speaking English. In addition, we should take the phonetic modeling of native-like English spoken by Korean people into account. To solve these problems, we introduced end-to-end ASR using an English database as an input and rule-based Konglish as an output. We expected the output of end-to-end ASR to make up for the shortcomings of the rules. Figure 3 shows the creative process of an enhanced Konglish DB.

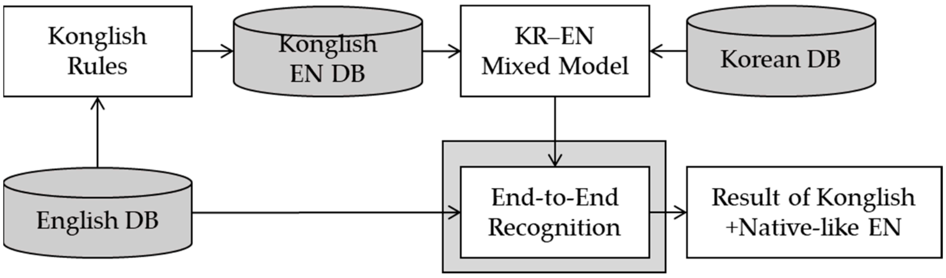


Figure 3. Flow diagram of the proposed method using an English database as training data. DB, database; EN, English; KR, Korean.

First, 1000 h of an English DB was converted into Konglish according to the rules. Second, we integrated these data with 1000 h of a Korean DB, and generated a mixed model through end-to-end ASR training. By an inferencing process based on the model, the rule-driven Konglish phoneme sequence can be enhanced with regard to CS pronunciation, as shown in the following examples:

1. English sentence (partial): Look, if there's anything I can do to make . . .
2. Rules only: 륵 입 데어즈 엔이싱 아이 캔 두 투 메익 . . . (Konglish phonemes: /lUk ip devzU enisiN ai kEn du tu meik/ . . .)
3. Hybrid (rules + deep learning): 룩 이프 데어스 애니싱 아이 캔 두 투 메이크 . . . (Konglish phonemes: /luk ipU devsU EnisiN ai kEn du tu meikU/ . . .)

Finally, 2000 h in total was prepared as a CS corpus.

4. LM Domain Adaptation Using Semantically Similar Sentences

We tried to cover the pronunciation variations and to balance the corpus in terms of AM. Still, due to the lack of CS data occurring in real life, we should take the infrequent occurrence of English words into account in terms of linguistic modeling. For that reason, we considered LM domain adaption using semantically similar sentences as the best way to approximate a target domain in real life. When similar sentences from a large text corpus were searched for, they had to include English words in which we were interested. Figure 4 shows the overall structure of the LM domain adaptation. As shown on the right side of the dotted line, the shallow fusion method incorporating a domain LM was used.

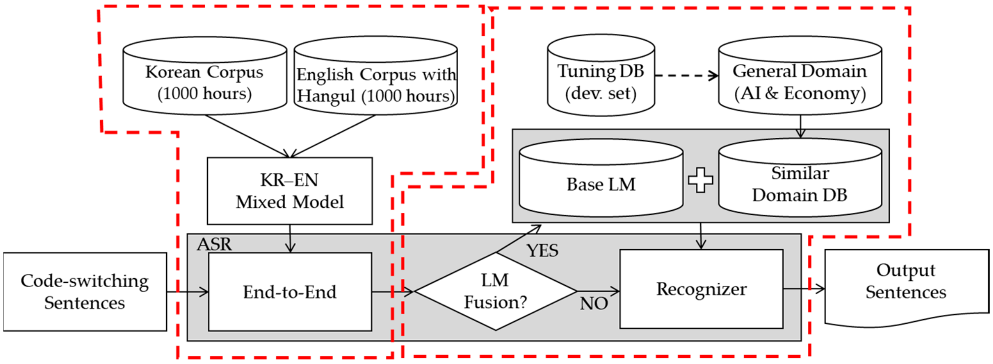


Figure 4. Overall structure of the Korean ASR for code-switching (CS); creating the Korean-Konglish mixed model (left dotted line); applying domain adaptation (right dotted line). ASR, automatic speech recognition; dev. set, development set; EN, English; KR, Korean; LM, language model.

The remaining problem was how to extract CS sentences that were semantically similar to the target domain. For this, it is reasonable to utilize a development set (dev. set) as a clue for the target domain. In this experiment, AI and economics lecture domains were chosen as the targets. The semantically similar sentences were extracted from the following three steps:

- Step 1: Sentences containing very rare English words (domain adaptation 1)

The domain adaptation DB consists of CS sentences containing very rare English words. In general, the lower the frequency, the less the ambiguity. For example, deep learning is almost definitely AI-related. Accordingly, sentences containing low-frequency English words should be included in a domain DB. Very rare English words can be found by counting occurrences from the general domain. The English words in the dev. set should then be compared to the very rare English words. If there is a match, the sentence is included in the domain DB. Figure 5 shows the steps in detail.

- Step 2: Sentences containing more than two English words (domain adaptation 2)

In general, if there are English words in a sentence, the topic of the sentence is likely close to the target domain. Hence, these sentences should be given preference for inclusion in a domain DB. For this reason, we extracted CS sentences that had many English words from the general domain text corpus. Duplicates of words in a sentence were not allowed. For example, one of the general domain sentences, "다이나믹 옵티마이저는 화질을 유지하면서 용량을 줄였다"—"the dynamic optimizer reduced the capacity while maintaining the image quality"—contains two foreign words: /*tainamik*/ (다이나믹; dynamic) and /*opthimaice*/ (옵티마이저; optimizer). This is used for generating a domain LM if these words are included in the dev. set. A total of 183,000 sentences were collected from the general domain text corpus containing words with two or more English words in the dev. set. Figure 6 shows the steps in detail.

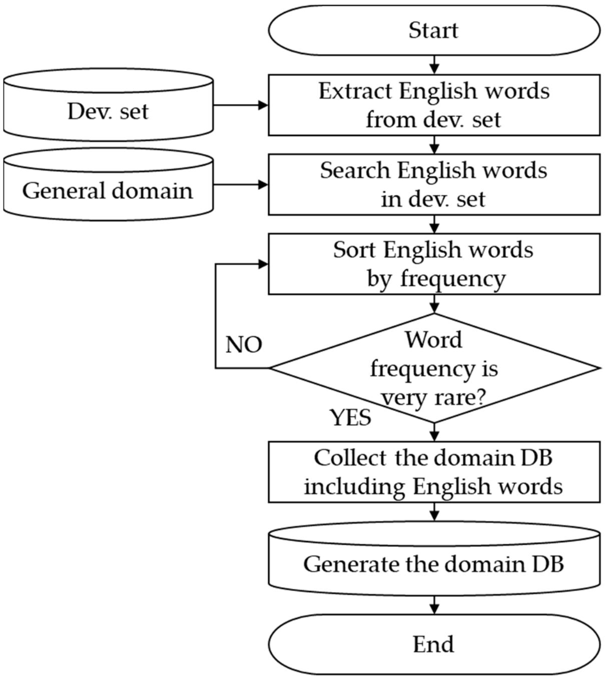

Start
Dev. set
Extract English words from dev. set
General domain
Search English words in dev. set
Sort English words by frequency
NO
Word frequency is very rare?
YES
Collect the domain DB including English words
Generate the domain DB
End

Figure 5. Block diagram of the sentences containing very rare English words. Dev. set, development set; DB, database.

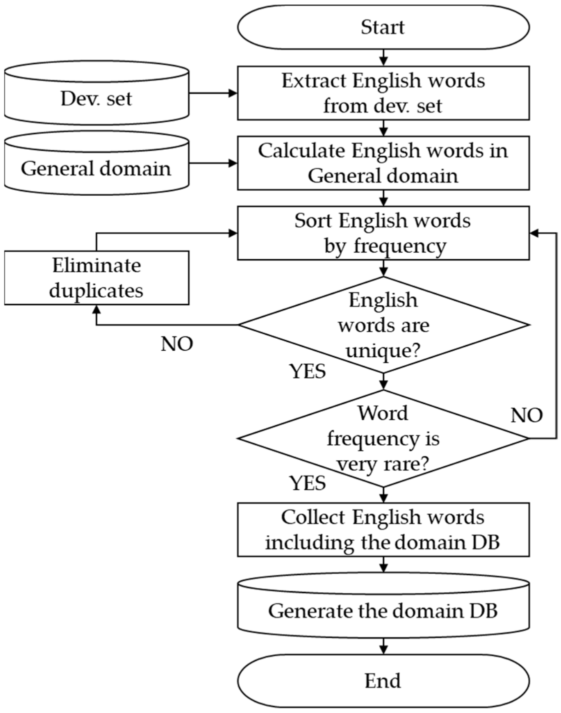

Start
Dev. set
Extract English words from dev. set
General domain
Calculate English words in General domain
Sort English words by frequency
Eliminate duplicates
English words are unique?
NO
YES
Word frequency is very rare?
NO
YES
Collect English words including the domain DB
Generate the domain DB
End

Figure 6. Block diagram of the sentences containing more than two English words. Dev. set, development set; DB, database.

- Step 3: Sentences semantically similar to the target domain (domain adaptation 3)

Recently, bidirectional encoder representations from transformers (BERT) [26] represented the semantic relationships of words in an embedding space well. In this study, we utilized KorBERT [27], which is specialized in Korean, and extracted CS sentences that were semantically similar to the dev. set from the general domain text corpus. Euclidian cosine similarity was adopted to measure the degree of similarity. The cosine similarity for arbitrary sentence vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as follows:

$$\cos\,(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}\cdot\boldsymbol{b}}{||\boldsymbol{a}||\;||\boldsymbol{b}||} \tag{1}$$

Figure 7 describes the steps with the cosine similarity.

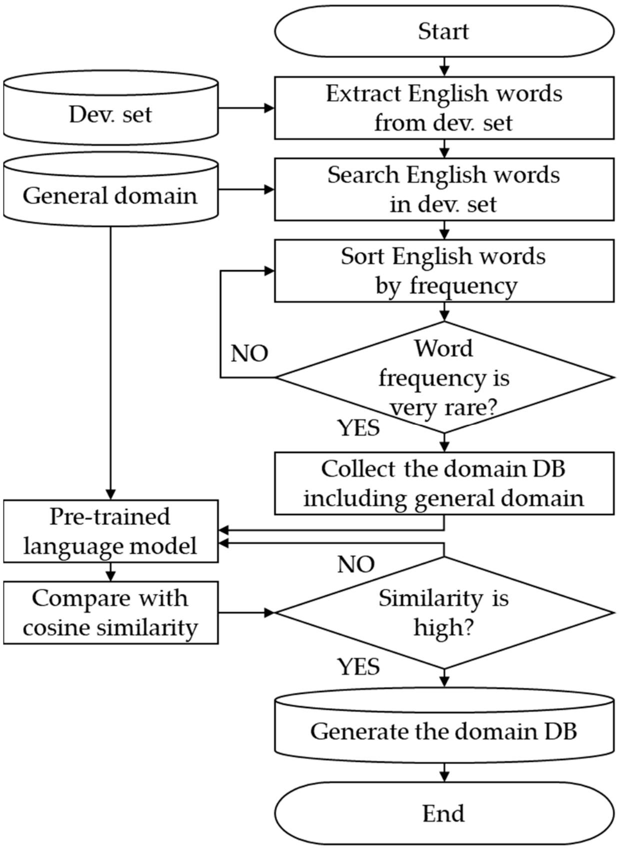

Start
Dev. set
Extract English words from dev. set
General domain
Search English words in dev. set
Sort English words by frequency
NO
Word frequency is very rare?
YES
Collect the domain DB including general domain
Pre-trained language model
NO
Compare with cosine similarity
Similarity is high?
YES
Generate the domain DB
End

Figure 7. Block diagram of the sentences semantically similar to the target domain. Dev. set, development set; DB, database.

The extracted sentences for the domain DB were converted into Konglish using the rules. Finally, the domain DB was combined with the base LM using shallow fusion [28], as shown in Equation (2):

$$\boldsymbol{y}^{*} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \log p(\boldsymbol{y}|\boldsymbol{x}) + \lambda_{\text{Base}} \log p_{\text{LM}_{\text{Base}}}(\boldsymbol{y}) + \lambda_{\text{Domain}} \log p_{\text{LM}_{\text{Domain}}}(\boldsymbol{y}) \tag{2}$$

where $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ is a sequence vector consisting of n elements, $\boldsymbol{y}$ is an existing output sequence vector, $\boldsymbol{y}^{*}$ is an expected output sequence vector, and λ_{Base} and λ_{Domain} are the weights of the base LM and the domain LM, respectively.

This idea can apply to English–Korean automatic speech translator in Figure 1. Using the proposed method, the recognition rate between English and Korean can improve via the model with LM domain adaptation.

5. Experimental Setup

5.1. Baseline System

The experiment was conducted on ESPnet [29], which supports a kind of end-to-end ASR framework. Our model is long short-term memory (LSTM)-based with listen, attend, and spell (LAS) [30] architecture, and it also uses a connectionist temporal classification

(CTC) hybrid model [31]. The input length was 600, and the output length was 150. The training data consisted of 1.06 million sentences (1052 h) of AI Hub Korean and 1.07 million sentences (960 h) of Librispeech [32] English. While learning, the training data composition was randomly mixed for each language. KR baseline represents the experiment trained using Korean alone, and KR–EN used both Korean and English. The wordpiece unit was adopted as a unigram subword model [33]. It is used for English–Chinese CS ASR study with end-to-end approach [34]. The output node was assigned to 3969 (i.e., 1950 nodes in Korean and 1946 nodes in English).

For evaluation, three types of test sets were prepared. The first, that is, the Economy set, consisted of 213 sentences from economics lectures in the business domain spoken by a Korean who can speak English like a native speaker. The second two, AILec1 (with 337 sentences) and AILec2 (with 623 sentences), comprised AI domain lectures spoken by a Korean speaker who could not speak English at all. In fact, AILec2 is a more difficult evaluation set than AILec1, since it has lots of explanations of mathematical expressions for formulas. There are 582 English words (19.1%) out of the 3047 words in the Economy set. In AILec1, there are 957 English words (20.8%) out of 4606 words, and in AILec2, there are 2211 English words (21.6%) out of 10,223 words.

5.2. Applying the Korean–Konglish Mixed Model

The proposed model, Konglish native-like EN, was trained using a Konglish DB (made from Librispeech) and AI Hub Korean DB. The experimental results of the KR baseline, KR–EN, and Konglish native-like EN, are shown in Table 3. The character accuracy (CA) of KR–EN is a little better than the baseline for Economy, but for AILec1 and AILec2, the CA was lower than that of the baseline. As expected, native English pronunciation of KR–EN was found for Economy, which was spoken by a Korean speaker with fluent English pronunciation. On the contrary, adding English data had a negative effect on recognition in both AILec1 and AILec2. Furthermore, the KR–EN in the latter two caused frequent confusion between English and Korean words at a rate of 2.8% of the total 213 sentences.

Table 3. Character accuracy of KR baseline, KR–EN, and Konglish native-like EN.

Character Accuracy (%)	Economy	AILec1	AILec2
KR Baseline	71.0	76.4	77.4
KR–EN [1]	72.6	76.3	76.7
Konglish Native-like EN [2]	74.3	77.8	78.5

[1] KR syllable and EN wordpiece. [2] Korean and Konglish mixed model.

Konglish native-like EN adopted syllable units, like the KR baseline, and the training DB was the same as KR–EN, except for the Konglish DB. The structure of the end-to-end ASR was almost the same as for KR–EN, but the output nodes were set to 1998 because of the reflection of unseen characters in Korean that English words transforming into the Hangul alphabet. As seen in Table 3, the performances for all test sets were improved meaningfully by integrating Konglish, which originated from native English.

5.3. Applying Domain Adaptation Using Shallow Fusion

LM domain adaptation is a process of shallowly fusing the base LM with a domain corpus. The base LM was trained based on recurrent neural network (RNN) LM using Korean and English text corpora. The dev. Set consisted of 168 Economy sentences and 1895 lecture sentences. Other parameters and hardware settings are described in Table 4. These are based on the hyperparameters of Librispeech, with some values adjusted.

Table 4. The configuration of the experimental setup for the model.

Configurations	Hyperparameters	
Model [1]	Encoder type	VGGBLSTM
	Encoder layers	5
	Decoder layers	2
	Encoder units	1024
	Projection units	1024
	Attention type	Location
	Attention dimension	1024
	Number of channels	10
	Filter size	100
Training	Optimizer	AdaDelta
	CTC weight	0.5
	Epochs	20
	Batch size	16
RNN LM [2] Training	Optimizer	SGD
	Epochs	185~660 [3]
Decoding	Beam size	10
	CTC weight	0.5
	LM weight	0.7

[1] Hardware is tested in Intel Xeon E5-2609 with 12 processors, and 128 GB memory per 1 node; in the experiment, we used 10 nodes. [2] An acronym of recurrent neural network language model. [3] Methods (Epochs): domain adaptation 1 (660); domain adaptation 2 (185); domain adaptation 3 (250).

The CS sentences for the domain adaption were extracted from 9.6 million sentences in the general domain text corpus. Among the English words sorted by frequency in the dev. Set, the CS sentences of each domain were extracted from the general domain corpus, including words with frequencies of 2000 or less. Figure 8 shows the frequency and the cumulative rate per English word, which consists of 380 ranks in the dev. set. This implies that the cumulative rate of the words remained almost above 95% in English word rank. The proportion of very rare English words in the word list was determined to be 95%, according to empirical experiments. Since very rare English words appear only in certain domains, they have the advantage of contributing to domain adaptation. Hence, it is regarded that a word frequency of 2000 or less is very rare for English words in Korean. Finally, both 13,708 and 23,395 CS sentences from each domain were extracted as the domain DBs.

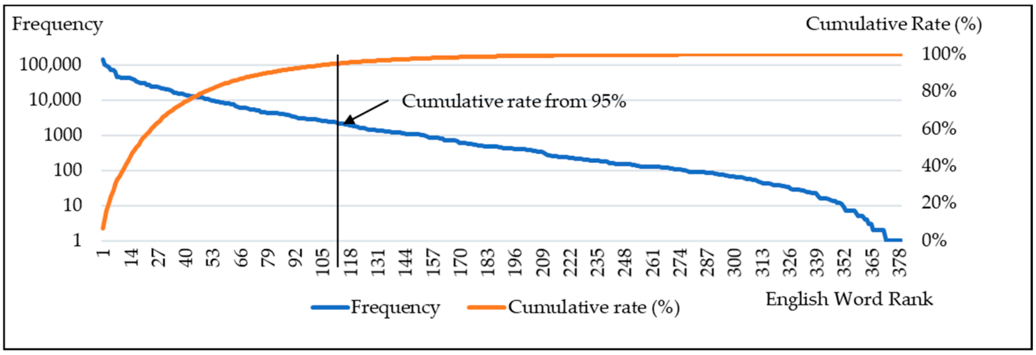


Figure 8. Finding very rare English words (domain adaptation 1); the frequency on a log scale for the left y-axis (dotted line); the cumulative rate (%) for the right y-axis (solid line); English word rank for the horizontal x-axis.

The threshold of cosine similarity was set to above 0.6. In this case, 21,218 and 86,786 sentences were found for each domain after removing duplications. For instance, Figure 9 describes the cosine similarity in ascending order of the dev. set of the lecture domain.

The word rank of the lecture domain which contains 108,000 words was chosen by the 107,995th to select the threshold value.

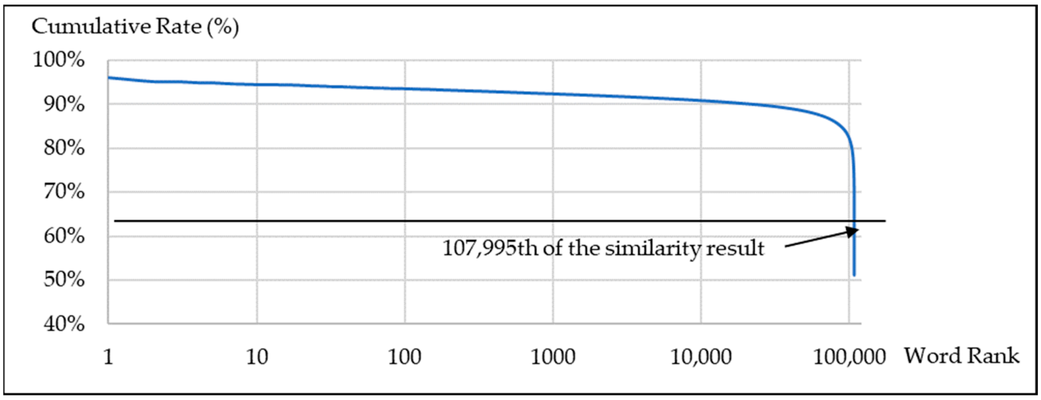


Figure 9. The cosine similarity in ascending order of the dev. set; the x-axis presents the number of sentences on a log scale; the y-axis shows the percentage of the cosine similarity.

Through steps 1–3 in Section 4, the domain DBs for adaptation can be summarized as shown in Table 5, where domain adaptation 1 is DA 1, domain adaptation 2 is DA 2, and domain adaptation 3 is DA 3. The base LM was trained with RNN LM using 2.1 million sentences in Korean and English. DA 1 + 2 + 3 means that all of the domain DBs were used for adaptation.

Table 5. The LM corpus for domain adaptation.

Model Name	Economy Dev. Set	Lecture Dev. Set
Base LM	2.1 million	2.1 million
Base LM + DA 1	Base LM + 13,708	Base LM + 23,395
Base LM + DA 2	Base LM + 3073	Base LM + 180,019
Base LM + DA 3	Base LM + 21,218	Base LM + 86,786
Base LM + DA 1 + 2 + 3 [1]	Base LM + 37,999	Base LM + 290,200
(Closed Test) [2]	Eco + AI1 + AI2 (1173) [3]	Eco + AI1 + AI2 (1173) [3]

[1] Sum of domain adaptations 1 (DA 1), 2 (DA 2), and 3 (DA 3). [2] The closed test was excluded from other comparison sets. [3] Sum of Economy, AILec1, and AILec2.

Table 6 shows the results of the LM adaptation. With the base LM only, the CA improved with Economy and AILec2, but not with AILec2. After adapting LM at each step (i.e., steps 1–3 in Section 4), the results show that the performance improved steadily when compared with the K-Base LM. When all of the domain DBs were combined (i.e., K-Base LM + DA 1 + 2 + 3), we obtained the highest performance among all of the combinations. The LM adaptation seemed to work well, as expected. Thus, the proposed method for selecting CS sentences is effective for ASR problems.

5.4. Analysis of English words

In this study, our intention was to improve the recognition of English words in Korean speech. We analyzed the recognition results and computed the accuracy of English words only. Table 7 shows the results.

The Konglish native-like EN with the K-Base LM + DA 1 + 2 + 3 method showed the best performance. The performance improved approximately 20% more than that of the KR baseline in all tasks.

Table 6. Character accuracy (CA) of the domain adaptation.

Set Names	CA (%)			ERR [1] (%)		
	Eco [2]	AI1 [2]	AI 2 [2]	Eco [2]	AI 1 [2]	AI 2 [2]
KR Baseline	71.0	76.4	77.4	-	-	-
Konglish Native-like EN	74.3	77.7	78.5	11.6	5.8	4.5
+Base LM (K-Base LM)	75.1	78.0	77.8	14.3	6.6	1.5
K-Base LM + DA 1	75.8	79.2	79.3	16.5	12.0	8.4
K-Base LM + DA 2	75.3	79.4	80.0	15.0	12.7	11.1
K-Base LM + DA 3	75.9	79.8	79.8	16.9	14.2	10.5
K-Base LM + DA 1 + 2 + 3	76.0	80.0	80.3	17.3	14.8	12.5
(Closed Test) [3]	(83.1)	(85.7)	(85.9)	-	-	-

[1] Error reduction rate (ERR) calculated by both KR baseline and another set (e.g., K-Base LM + DA 2 or K-Base LM + DA 3). [2] Eco is an acronym of Economy; AI1 is known as AILec1; AI2 is the short form of AILec2. [3] Used the shallow fusion method as the closed test; it was excluded from the other sets.

Table 7. ASR correction rate of English words with Hangul in Korean and English CS text.

Correction Rate (%)	Economy	AILec1	AILec2
KR Baseline	32.7	42.0	38.3
Konglish Native-like EN	42.9	44.4	44.0
+Base LM (K-Base LM)	46.7	50.4	46.8
K-Base LM + DA 1	51.2	58.6	61.4
K-Base LM + DA 2	51.9	63.3	63.8
K-Base LM + DA 3	52.1	59.5	61.1
K-Base LM + DA 1 + 2 + 3 [1]	54.0	60.7	62.5

[1] Sum of domain adaptations 1, 2, and 3.

6. Conclusions

In the case of Korean, English words pronounced by Korean speakers—Korean-style English (i.e., Konglish)—have many phonetic variations from native-like English pronunciation. Moreover, mixed-language spoken data are very rare, making any model biased toward Korean, even if there are many data. Pronunciation variations and imbalanced data are major problems that degrade the recognition of CS speech.

In this paper, we proposed pronunciation variations reflecting English words spoken by Koreans and the LM adaptation based on similarity of meaning. First, we tried to find a unified pronunciation model based on phonetic knowledge and deep learning by applying the language identification (LID) of Watanabe et al. [4]. Despite this, there were problems with intrusions occurring between languages. However, our proposed method can avoid this problem.

Secondly, we extracted the CS sentences that were semantically similar to the target domain and then applied the language model (LM) adaptation to solve the biased modeling toward Korean due to the imbalanced training data. Nakayama et al. [18] utilized a speech chain framework based on deep learning to enable ASR and TTS to learn code-switching. Although this closed-loop architecture improves the performance even without any parallel code-switching data, there is a limit to improving the performance when only using synthetic speech due to the quality. It seems that the performance can be improved if this method is combined with our method.

Compared with the KR baseline, the proposed hybrid method (e.g., knowledge and deep learning) showed up to 11.6% improvement in the error reduction rate (ERR). Through the semantically similar sentence extraction process, we were able to obtain 16.5%, 15.0%, and 16.9% improvements in ERR in the experiments of LM adaptation. If all domain DBs were combined, the ERR improved by up to 17.3%. LM adaptation using the proposed method might be one way to solve the biased data problem, which is critical.

However, although we dealt with some critical issues, if compared to the closed test, which would be the upper bound of the performance, there is still room for improvement.

Recently, Tacotron [35], which is a kind of text-to-speech (TTS) system, produced very high-quality synthesized speech. Thus, Tacotron should be incorporated into our model to cope with the CS problem. Additionally, cross-lingual speech and the text embedding method will be helpful to improve the performance of our model.

Author Contributions: Conceptualization, D.L. and S.K.; methodology, D.L.; software, D.L. and D.K.; validation, D.L. and S.K.; writing—original draft preparation, D.L.; writing—review and editing, D.K., S.Y., and S.K., supervision, S.K.; project administration, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by an Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. (21ZS1140, Conversational AI Core Technology Research).

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to the use of the speech database of human voice recognition.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in a publicly accessible repository (AI Hub and Librispeech).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zirker, K.A.H. Intrasentential vs. Intersentential Code Switching in Early and Late Bilinguals. Master's Thesis, Bringham Young University, Bringham, UK, 2007.
2. Yeong, Y.-L.; Tan, T.-P. Applying Grapheme, Word, and Syllable Information for Language Identification in Code Switching Sentences. In Proceedings of the 2011 International Conference on Asian Language Processing, Penang, Malaysia, 15–17 November 2011; pp. 111–114.
3. Chan, J.Y.C.; Ching, P.C.; Lee, T.; Meng, H.M. Detection of Language Boundary in Code-Switching Utterances by Bi-Phone Probabilities. In Proceedings of the 2004 International Symposium on Chinese Spoken Language Processing, Hong Kong, China, 15–18 December 2004; pp. 293–296.
4. Watanabe, S.; Hori, T.; Hershey, J.R. Language Independent End-to-End Architecture for Joint Language Identification and Speech Recognition. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop, Okinawa, Japan, 16–17 December 2017; pp. 265–271.
5. Seki, H.; Hori, T.; Watanabe, S.; Le Roux, J.; Hershey, J.R. End-to-End Multilingual Multi-Speaker Speech Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 3755–3759.
6. Lee, G.; Ho, T.-N.; Chng, E.-S.; Li, H. A Review of the Mandarin-English Code-Switching Corpus: SEAME. In Proceedings of the 2017 International Conference on Asian Language Processing, Singapore, 5–7 December 2017; pp. 210–213.
7. Yılmaz, E.; McLaren, M.; van den Heuvel, H.; van Leeuwen, D.A. Semi-supervised acoustic model training for speech with code-switching. *Speech Commun.* **2018**, *105*, 12–22. [CrossRef]
8. Yue, X.; Lee, G.; Yılmaz, E.; Deng, F.; Li, H. End-to-End Code-Switching ASR for Low-Resourced Language Pairs. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop, Singapore, 14–18 December 2019; pp. 972–979. [CrossRef]
9. Hussain, A.; Arshad, M.U. An Attention Based Neural Network for Code Switching Detection: English & Roman Urdu. *arXiv* **2021**, arXiv:2103.02252.
10. Li, Y.; Fung, P. Improved Mixed Language Speech Recognition Using Asymmetric Acoustic Model and Language Model with Code-Switch Inversion Constraints. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 16–31 May 2013; pp. 7368–7372.
11. 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, Kyoto, Japan, 25–30 March 2012; pp. 4889–4892.
12. Pandey, A.; Srivastava, B.M.L.; Gangashetty, S.V. Adapting Monolingual Resources for Code-Mixed Hindi-English Speech Recognition. In Proceedings of the 2017 International Conference on Asian Language Processing, Singapore, 5–7 December 2017; pp. 218–221.
13. Emond, J.; Ramabhadran, B.; Roark, B.; Moreno, P.; Ma, M. Transliteration Based Approaches to Improve Code-Switched Speech Recognition Performance. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop, Athens, Greece, 18–21 December 2018; pp. 448–455.
14. Chandu, K.R.; Black, A.W. Style Variation as a Vantage Point for Code-Switching. *arXiv* **2020**, arXiv:2005.00458.

15. Long, Y.; Li, Y.; Zhang, Q.; Wei, S.; Ye, H.; Yang, J. Acoustic data augmentation for Mandarin-English code-switching speech recognition. *Appl. Acoust.* **2020**, *161*. [CrossRef]
16. Chang, C.-T.; Chuang, S.-P.; Lee, H.-Y. Code-Switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation. In Proceedings of the 2019 Interspeech, Graz, Austria, 15–19 September 2019; pp. 554–558.
17. Tjandra, A.; Sakti, S.; Nakamura, S. Listening While Speaking: Speech Chain by Deep Learning. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop, Okinawa, Japan, 16–17 December 2017; pp. 301–308.
18. 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 2018 IEEE Spoken Language Technology Workshop, Athens, Greece, 18–21 December 2018; pp. 182–189.
19. Nakayama, S.; Tjandra, A.; Sakti, S.; Nakamura, S. Zero-Shot Code-Switching ASR and TTS with Multilingual Machine Speech Chain. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop, Singapore, 14–18 December 2019; pp. 964–971.
20. Yu Cho, Y. Korean Phonetics and Phonology. *ORE Linguist.* **2016**. [CrossRef]
21. Kent, D.B. Speaking in Tongues: Chinglish, Japlish and Konglish. In Proceedings of the Korea TESOL Proceedings the Second Pan Asian Conference, Seoul, Korea, 1–3 October 1999.
22. Cho, B.-E. Issues concerning Korean Learners of English: English Education in Korea and Some Common Difficulties of Korean Students. *East Asian Learn.* **2004**, *1*, 31–36.
23. CMUdict. Available online: http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/ (accessed on 7 October 2020).
24. Lee, D.; Yun, S.; Kim, J.; Kim, S. A Prelimanary Study on the Pronunciation Rule of Korean Pronoun for English Speakers. In Proceedings of the 2017 Oriental COCOSDA International Conference on Speech Database and Assessments, Seoul, Korea, 1–7 November 2017; pp. 210–213.
25. Novak, J.R.; Minematsu, N.; Hirose, K. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Nat. Lang. Eng.* **2016**, *22*, 907–938. [CrossRef]
26. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
27. KorBERT. Available online: http://aiopen.etri.re.kr/service_dataset.php (accessed on 7 October 2020).
28. Gülçehre, Ç.; Firat, O.; Xu, K.; Cho, K.; Barrault, L.; Lin, H.-C.; Bougares, F.; Schwenk, H.; Bengio, Y. On Using Monolingual Corpora in Neural Machine Translation. *arXiv* **2015**, arXiv:1503.03535.
29. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech Processing Toolkit. *arXiv* **2018**, arXiv:1804.00015.
30. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 4960–4964.
31. Hori, T.; Watanabe, S.; Hershey, J. Joint CTC/Attention Decoding for End-to-End Speech Recognition. In Proceedings of the 2017 Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 518–529.
32. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
33. Kudo, T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In Proceedings of the 2018 Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 66–75.
34. Zeng, Z.; Khassanov, Y.; Pham, V.T.; Xu, H.; Chng, E.S.; Li, H. On the End-to-End Solution to Mandarin-English Code-switching Speech Recognition. *arXiv* **2018**, arXiv:1811.00241.
35. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards End-to-End Speech Synthesis. *arXiv* **2017**, arXiv:1703.10135.