electronics

Article

Automatic Speech Recognition of Vietnamese for a New Large-Scale Corpus

Linh Thi Thuc Tran [1,*], Han-Gyu Kim [2], Hoang Minh La [1] and Su Van Pham [1]

[1] Faculty of Electronics Engineering, Posts and Telecommunications Institute of Technology, Hanoi 100000, Vietnam; hoanglm.b19dt093@stu.ptit.edu.vn (H.M.L.); supv@ptit.edu.vn (S.V.P.)
[2] NAVER Cloud, Seongnam 13529, Republic of Korea; hangyu.kim@navercorp.com
* Correspondence: linhtt@ptit.edu.vn

Abstract: Vietnamese is an under-resourced language. The requirement for a large-scale and high-quality Vietnamese speech corpus increases on demand. We introduce a new large-scale Vietnamese speech corpus with 100.5 h collected from various audio sources in the Internet. The raw collected audio was processed to obtain clean speech. Transcription of the clean speech was made manually. The new corpus was analyzed in terms of gender, topic and regional dialect. Results shows that the new corpus has good diversity of genders, topics and regional dialects. We also evaluated the new corpus using state-of-the-art automatic speech recognition models like LAS and Speech-Transformer for multiple scenarios. This is the first time that these models have been applied to Vietnamese speech recognition and obtained reasonable results. Simulation results showed that the new corpus would be a good dataset for the Vietnamese ASR tasks because it reflected correctly difficulties in recognizing speech from different dialects and topic domains.

Keywords: Vietnamese corpora; automatic speech recognition; LAS; transformer; SpecAugment

check for updates

Citation: Tran, L.T.T.; Kim, H.-G.; La, H.M.; Van Pham, S. Automatic Speech Recognition of Vietnamese for a New Large-Scale Corpus. *Electronics* **2024**, *13*, 977. https://doi.org/10.3390/electronics13050977

Academic Editors: Junhua Ding, Haihua Chen, Yunhe Feng and Tozammel Hossain

Received: 31 December 2023
Revised: 23 February 2024
Accepted: 1 March 2024
Published: 4 March 2024

1. Introduction

Vietnamese is spoken by approximately 90 million speakers, primarily in Vietnam. It serves as the official language for the majority of the country's population. The distinctiveness of Vietnamese is that it is a monosyllabic and tonal language. Moreover, the Vietnamese writing script is based on the Latin script with additional diacritics to represent tones, consisting of 91 characters in total. Vietnamese belongs to under-resourced languages since the Vietnamese resources that are publicly available are very limited. Additionally, most public corpora of Vietnamese are constructed by recording speakers' voices when they read given texts [1–3]. This means that the available datasets might not be compatible for realistic scenarios relating to spoken language, such as conversational and discussion recognition [4]. Some Vietnamese speech corpora are generated by crawling speech from the Internet or by combining reading text and crawling speech. Many current Vietnamese corpora have a small size, around a few hours to tens of hours, such as VIVOS, VLSP 2018, etc., [2,3,5]. The Vietnamese corpora of more than 100 h, such as VinBigdata-VLSP2020 and corpora in [6,7], are rare, and most of them are not open-access, like the corpus collected by FPT Technology Research Institute (FTRI), namely FTRI corpus, MICA VNSpeechCorpus [8], and the corpora in [6,7]. However, those corpora are not either high-quality sound or open-access. Note that constructing a large-scale and high-quality corpus costs a lot of time and effort. The demand for large-scale and high-quality Vietnamese corpora that reflect real-life situations has increased for research purposes.

In this paper, we introduce a new large-scale corpus of Vietnamese with total of 100.5 h of nearly clean speech. This corpus is constructed by crawling many audio resources in the Internet and manually transcribing them. Unlike other Vietnamese corpora, the audio files in our collected corpus are labeled not only by gender and regional dialect but also by various topics such as news, reading, healthcare, tourism, sports, etc. Our new corpus is open-access and can be found at the link: https://drive.google.com/drive/folders/1tiPKaIOC7bt6isv5qFqf61O_2jFK8ZOI?usp=sharing, accessed on 1 March 2024.

Most previous works for Vietnamese automatic speech recognition (ASR) tasks investigated traditional statistical ASR models like HMM/GMM, deep neural networks (DNNs) or hybrid models—a combination of HMM/GMM and DNN models [3,6,7,9–11]. Recently, end-to-end (E2E) models in the field of automatic speech recognition have gained considerable attention from both academic and industrial perspectives [12–17]. These models integrate the components of traditional ASR systems—specifically, the acoustic model (AM), pronunciation model (PM) and language model (LM)—into a unified neural network. The optimization of these components is performed jointly within the E2E model. This integration results in a simplified system architecture, facilitating ease of development and maintenance. Among them, Listen, Attend and Spell (LAS) [18], Transformer [16] and Speech-Transformer (sTransformer) [19] are state-of-the-art (SOTA) E2E models. To the best of our knowledge, those SOTA approaches have been used for ASR on some rich-resourced languages such as English and Mandarin, but not for Vietnamese, a low-resourced language with monosyllables and six tones. Therefore, we propose to apply LAS, Transformer and their variants for ASR tasks to evaluate the effectiveness of the collected corpus in multiple scenarios. Simulation results showed that the behaviour of the new corpus was consistent, and it would be a good dataset for the Vietnamese ASR tasks.

The contribution of this paper is twofold:

(i) To introduce a novel large-scale Vietnamese speech corpus (LSVSC);
(ii) To propose to use SOTA end-to-end automatic speech recognition approaches such as LAS and sTransformer with/without the integration of advanced techniques (e.g., SpecAugment (SA) and adaptive SpecAugment (adaptSA)) on Vietnamese.

This paper is organized as follows. Section 2.1 describes the main characteristics of the Vietnamese language. A review of previous Vietnamese speech corpora and ASR methods has been mentioned in Section 3. Section 4 introduces a new large-scale Vietnamese speech corpus that we collected. Section 5 represents the application of the SOTA end-to-end ASR models such as LAS, sTransformer and their variants on the new corpus. Simulation results are presented in Section 6, where we evaluate the LSVSC by using SOTA end-to-end ASR models with multiple scenarios. Section 7 discusses the results and highlights future work.

2. Vietnamese Language

2.1. Characteristics of Vietnamese Language

Vietnamese is the official and national language of Vietnam, with about 90 million native speakers. It serves as the primary language for the majority of the country's population, and it is also spoken as a first or second language by various ethnic minority groups within Vietnam. Indeed, the Vietnamese language is known for its unique characteristics that make it quite distinctive from many other languages. In particular, Vietnamese words are often monosyllabic, i.e., each word consists of one syllable. Additionally, Vietnamese is a tonal language, which means that the pitch or intonation used when pronouncing a word can change its meaning. There are six tones in Vietnamese, each represented by a specific diacritical mark or indicated by the absence of a mark, cf. Table 1. The differences among tones in Vietnamese are illustrated in Table 2 [7]. The monosyllabic nature of Vietnamese, combined with its tonal system, adds a layer of complexity to the language.

Table 3 shows the structure of Vietnamese syllables. The Vietnamese language has 22 initial phonemes and 16 final phonemes. Initials are the consonant sounds that begin a syllable, while finals are the vowel sounds that follow the initial. While there is a total of 19,000 unique possible syllables in Vietnamese, only around 6500 of them are used in practical communication. This highlights the richness and variety of potential combinations in the language.

The Vietnamese writing system is based on the Latin alphabet with additional diacritics to represent tones, making it distinct from many other languages in the region that use different scripts.

Table 1. Tones in Vietnamese.

Tones	**Description**
Low-Falling Tone	Indicated by a grave accent
High-Broken Tone	Indicated by a hook above the letter
Low-Rising Tone	Indicated by a tilde
High-Rising Tone	Indicated by an acute accent
Low-Broken Tone	Indicated by a dot below the letter
Mid Tone	No diacritical mark

Table 2. Structure of Vietnamese tones.

Pitch Contour	**Flat**	**Unflat**	
		Broken	**Unbroken**
High	No mark	High-broken	High-rising
Low	Low-falling	Low-rising	Low-broken

Table 3. Structure of Vietnamese syllables.

TONE			
Initial	FINAL		
	Onset	Nucleus	Coda

2.2. Challenges in Developing Vietnamese ASR Systems

Developing Vietnamese ASR systems copes with some challenges [4]. Specifically, the regional variations in pronunciation, vocabulary, and dialects in Vietnamese pose a significant challenge for automatic speech recognition (ASR) systems. There are three main regional dialects—Northern, Central and Southern—and each have distinct features that can affect the way words are spoken and understood.

Here are some examples of regional differences:

Northern Dialect:

The pronunciation of "l" as "n" and vice versa is a notable characteristic in some Northern regions. In addition, there are distinctive vocabulary and intonation patterns compared to other regions.

Central Dialect:

Features that differentiate this from the Northern and Southern dialects include unique vocabulary and pronunciation patterns.

Southern Dialect:

The pronunciation of ending consonants in "n" and "ng" may be similar in some Southern regions. Moreover, there are vocabulary differences and unique intonation.

These regional variations can make it challenging for ASR systems to accurately transcribe speech, especially when they are trained on data from one specific region but are exposed to speech from another. The diversity in dialects and pronunciation within each region further adds complexity.

Moreover, the variation in speaking style and the usage of different words to express the same meaning in different regions are common challenges in natural language processing, including the development of language models and automatic speech recognition (ASR) systems. Recording environments also play a crucial role in the performance of speech recognition systems. Different recording environments introduce various types and levels of noise, which can significantly impact the accuracy of the recognition results.

To address these challenges, ASR systems for Vietnamese may need to be trained on diverse datasets that represent the linguistic diversity across regions and are robust and adaptable to various speaking styles. Furthermore, training speech recognition models on dataset that include a variety of recording environments to make the system more adaptable to different acoustic conditions.

3. Previous Works

3.1. Previous Works on Vietnamese Speech Corpus

In this part, we conduct a survey about the Vietnamese speech corpora, cf. Table 4. In [2], a Vietnamese corpus of reading speak was constructed by asking native speakers from Hanoi and Ho Chi Minh City in Vietnam and 20 native speakers living in Karlsruhe, Germany to read prompted sentences extracting from Vietnamese e-newspapers. As a result, a speech corpus of 25 h spoken by 90 male speakers and 70 female speakers was collected. VIVOS [3], an open-access Vietnamese speech corpus, was released by AILAB of VNUHCM—University of Science (Vietnam) in 2017. It consists of 15 h of recording speech with 12,420 utterances prepared for the ASR task. There are 65 speakers, among them 34 are male and 31 are female speakers. In the same year, Viettel group—a Vietnamese corporation of multinational telecommunications and technology—collected 85.8 h of phone calls from the Viettel customer service call center [9]. The data were sampled by 8 kHz, with a resolution of 8 bits/sample.

Table 4. Summary of Vietnamese speech corpora.

Corpus	Size	Style	Open/Close
Corpus in [2]	25 h	Reading	Close
VIVOS [3]	15 h	Reading	Open
Viettel corpus [9]	85.8 h	Phone call	Close
MICA VNSpeechCorpus [8]	100 h	Reading	Close
FTRI corpus	2036 h	Reading	Close
Corpora in [7]: 2 small corpora 1 large-scale corpus	6 and 6.5 h 900 h	Reading Spontaneous	Close Close
VinBigdata-VLSP2020	20 h 80 h	Reading Spontaneous	Open Open

MICA VNSpeechCorpus [8], a large-scale corpus of reading speech, was constructed in 2005. It contains about 100 h of audio recordings from 50 native speakers. Firstly, the Vietnamese text corpus was prepared by collecting texts from various resources in the Internet. Secondly, sentences from the text corpus were read by male and female speakers. The recordings were carried out in both quiet and office environments. Another large-scale corpus of Vietnamese speech with a total duration of 2036 h was collected by the FPT Technology Research Institute (FTRI). The FTRI corpus consists of reading speech with sentences extracted from daily news and forum websites. The corpus was constructed by recording voice of 3059 male and female speakers. The speakers represented the Northern, Central, and Southern dialects of Vietnam. All audio files were converted to the wave format with a sampling frequency of 16 kHz and PCM 16 bits. In [7], three Vietnamese speech corpora have been introduced. Those corpora include two small reading speech corpora with a total of 6 h and 6.5 h, respectively, and a large-scale speech corpus with 900 h. The large-scale speech corpus was collected by crawling untranscripted audio from various resources, such as movies, YouTube movies, and electronic newspapers. Then, the audio was stored in the format of PCM 16 bits, a sampling rate of 16 Khz and a mono channel. To transcribe such a large number of audio files, a hybrid text transcription consisting of an ASR system followed by manual verification and revision was employed. A large-scale Vietnamese speech corpus that is open-access is VinBigdata-VLSP2020, released in 2020. This corpus consists of approximately 100 h of speech. Among them there are approximately 20 h of reading speech and 80 h of spontaneous speech. The reading speech was recorded with a smartphone in various environments, while the spontaneous speech was crawled from the Internet and manually transcribed. The corpus was developed for the ASR task in VLSP-2020.

In the past twenty years, there have been a lot of efforts to increase the number of Vietnamese speech corpora, but the number of large-scale Vietnamese speech corpora has been limited. Most public corpora of Vietnamese speech are reading speech, like the FTRI corpus and the corpora in [2,3,8], or part of reading speech, such as VinBigdata-VLSP2020 or two small corpora in [7]. Therefore, the available corpora might not be compatible with real-life scenarios for spoken language, like conversational and discussion recognition. Moreover, most Vietnamese speech corpora are not either free-access or high-quality.

3.2. Previous Works on Vietnamese ASR

Traditionally, the components of speech recognition systems include acoustic, pronunciation, and language models. Those models were trained separately, each with its specific objective. In the past, hidden Markov models (HMM) and Gaussian mixture models (GMM) were commonly used for the acoustic model. These models are statistical in nature and have been applied to capture the relationships between acoustic features and phonemes. With the rise of deep learning, deep neural networks (DNNs) have become a popular choice for the acoustic model in recent studies [6,7]. DNNs are capable of learning complex patterns and representations from large amounts of data, making them well suited for acoustic modeling in ASR. Some studies have explored hybrid models [3,9,10], combining elements of both traditional statistical models (HMM/GMM) and DNNs. This combination leverages the strengths of both approaches. DNNs also have shown significant improvements in pronunciation models, particularly in mapping words to phoneme sequences. This is crucial for accurate speech recognition, since it enhances the model's ability to understand and represent spoken language.

Many studies have traditionally used n-gram models for language modeling [3,6,7,9,10]. N-gram models estimate the probability of a word based on the context of the preceding n-1 words. While effective, these models have limitations in capturing long-range dependencies. Some recent studies have explored the use of DNNs for language modeling. DNN-based language models have advantage of capturing more complex relationships in the data and performing well on tasks requiring context understanding. Recurrent models, a type of neural network architecture designed to handle sequential data, have been employed in language modeling. These models improve speech recognition accuracy by rescoring n-best lists, contributing to more effective transcription.

Overall, the combination of acoustic, pronunciation and language models, along with advancements in deep learning techniques, contributes to the continuous improvement of ASR systems for the Vietnamese language.

Recent advancements focus on addressing the disjoint training issue by adopting end-to-end training approaches [12,20–22]. In those approaches the functions of traditional ASR components such as acoustic, pronunciation and language models are combined into a single neural network. This means training models directly from speech to transcripts, aiming for a more integrated and streamlined learning process. Two main approaches for end-to-end training are highlighted: connectionist temporal classification (CTC) [20] and sequence-to-sequence models with attention [18,23,24]. The CTC model is designed to handle sequences of variable lengths and assumes conditional independence of label outputs. The sequence-to-sequence with attention model has been successfully applied to phoneme sequences, and trained end-to-end for speech recognition [17]. Attention mechanisms allow the model to focus on relevant parts of the input sequence when generating the output.

4. New Large-Scale Vietnamese Speech Corpus

In this section, we introduce a new large-scale Vietnamese speech corpus (LSVSC) crawled from various open sources in the Internet. We used Audacity software to extract a single channel of the recordings and remove all noisy segments from the raw data, and only kept the nearly clean segments. In fact, to remove noisy segments, the guideline-trained crowd workers played raw audio files, listened to recognize segments with audible noise, marked those segments and then removed them. The rest contained only audio segments without perceivable noise levels or with very low noise levels. We used the noise reduction function of Audacity to further reduce noise for audio segments with very low noise levels. As a result, we obtained nearly clean speech segments, which then were chopped into shorter segments with a length not more than 13 s, corresponding to a simple or a short-complexed sentence. The clean segments were saved, converted to a wave format with PCM 16 bits and resampled by a sampling frequency of 16 kHz. Those clean speech segments were double-checked by another group of the guideline-trained crowd workers to ensure that all speech segments with perceivable noise had been removed. We chose a manual way based on the subjective evaluation to prepare the corpus. Although this way costs time and human effort, the high quality of the speech signal can be achieved.

The clean corpus consists of all audio files after being preprocessed as described above. Specifically, the LSVSC consists of 100 h and 30 min of clean utterances, which covers various topics such as news, reading, audio books, movies, sports, healthcare, traffic, tourism, etc. Figure 1 illustrates the distribution of the LSVSC according to ten topics. News is the dominant topic, comprising 78.5% of the total dataset. The topic of reading, storytelling and audiobooks accounts for 9.41%. Health and Daily life topics represent 4.71% and 3.66%, respectively. Other topics collectively make up less than 1% each, with tourism being the least at about 0.4%. The clean LSVSC is manually transcribed by a group of crowd workers and double-checked by another group of crowd workers to ensure that the errors in the transcription are corrected. Finally, the text is saved in file as the UTF-8 format.

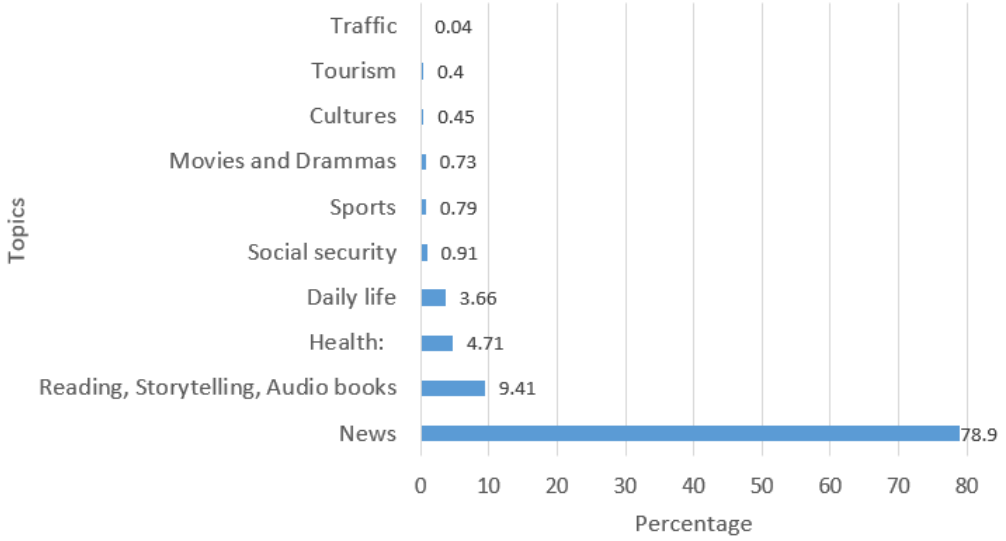


Figure 1. Voice distribution by topics in the collected clean corpus.

Figure 2 shows the voice distribution by gender in the LSVSC. The number of sentences utterred by female and by male speakers are approximately equal. Figure 3 describes the voice distribution by regional dialects in the LSVSC. The majority of the data in the corpus are spoken by the Northern dialect speakers because the Northern dialect is the standard dialect in Vietnam and very popular in the media. Central and Southern dialects account for 7.65% and 3.54%, respectively. The Central highland dialect represents 0.66%, and the minority ethnic group dialect is only 0.05%. The voice distribution by the length of utterances is illustrated in Figure 4. In this corpus, most utterances have the length from 3 s to 10 s. The longest utterances have a length of 13 s, whereas the length of the shortest ones is about 1 s.

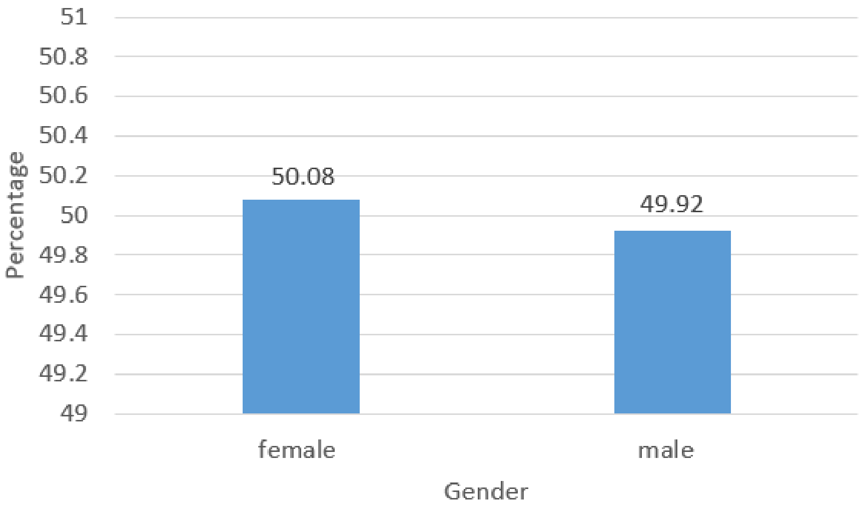


Figure 2. Voice distribution by gender in the collected clean corpus.

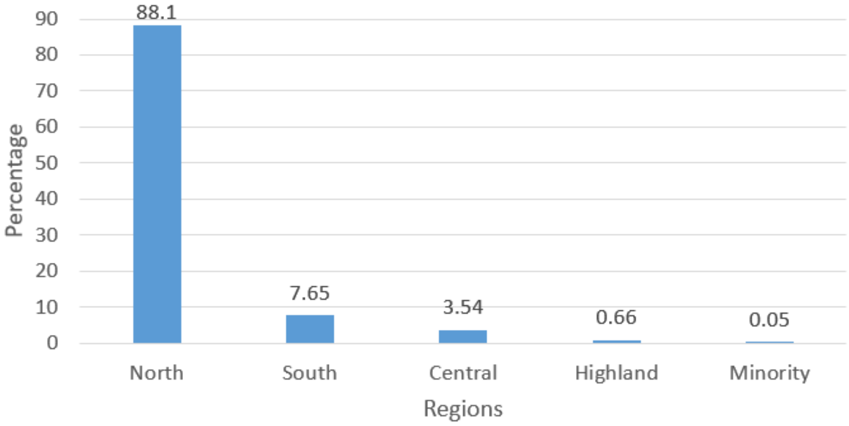


Figure 3. Voice distribution by regional dialects in the collected clean corpus.

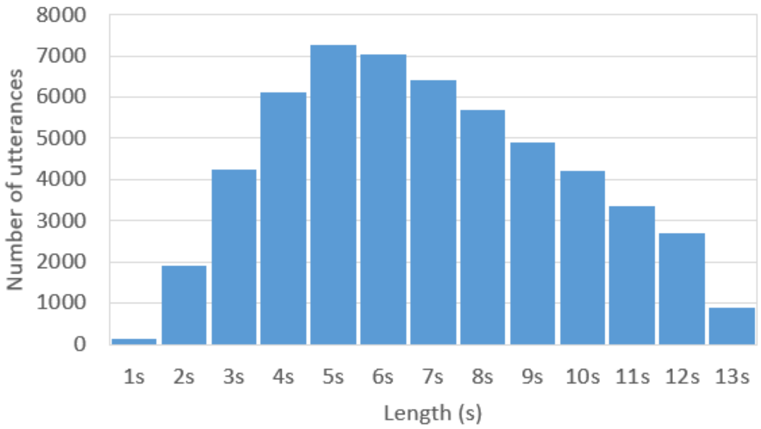


Figure 4. Voice distribution by length in the collected clean corpus.

5. End-to-End Vietnamese ASR Models

In this section, we propose to apply E2E state-of-the-art ASR models such as LAS, Speech-Transformer and their variants to the LSVSC. It is the first time that these end-to-end ASR models have been applied to the Vietnamese speech and showed reasonable performance. We train, evaluate and compare those models in terms of character error rate (CER) and word error rate (WER). Note that LAS and Transformer have been used for ASR on some rich-resourced languages such as English and Mandarin, but not for Vietnamese, a low-resourced language with monosyllables and six tones. We also generated a vocabulary for the LSVSC by scanning the transcription (text) corresponding to each audio sample in the corpus. Then, each text was separated into words with a space between two words. We compared those words to the words available in the vocabulary. If a word was not shown in the vocabulary, it was added to the vocabulary. We repeated the process until reaching the end of the corpus. Finally, we obtained a vocabulary of 6740 non-overlapping words. The details of the above models, their setups and performance are described in the following.

5.1. LAS Model

To address the limitations of CTC and sequence-to-sequence models with attention in the field of speech recognition, the Listen, Attend and Spell (LAS) neural network has been introduced in [18]. The LAS model is designed to transcribe an audio sequence signal into a word sequence character by character. It is based on the sequence-to-sequence learning framework with attention. LAS consists of an encoder recurrent neural network (RNN) and a decoder RNN. The encoder RNN, called the listener, is described as a bidirectional long short-term Memory RNN (BLSTM) with pyramidal structure. Its role is to convert low-level speech signals into higher-level features, capturing hierarchical representations. The decoder RNN (e.g., LSTM), called the speller, converts the higher-level features generated by the listener into output utterances. It achieves this by specifying a probability distribution over sequences of characters using the attention mechanism. The listener and the speller are trained jointly, i.e., both components of the LAS network are optimized simultaneously during the training process. In LAS, the elimination of independence assumptions and the use of joint training contribute to a more integrated and effective approach for speech recognition. This type of model is particularly well-suited for E2E training, where the entire system is optimized for the specific task without relying on intermediate components like HMMs.

In this work, we implemented LAS and LAS with SpecAugment models to recognize Vietnamese speech automatically. We trained and evaluated those models on the LSVSC. For the LAS model, we chose the number of encoders = 4, the number of decoders = 2, dropout = 0.3; the size of both the encoder and decoder were 512. The sampling rate, window size, window stride were 16 kHz, 400 samples and 160 samples, respectively. The models were trained with an initial learning rate of 0.0003 and utilized the Linear Warmup Scheduler with 7105 warm-up steps. We chose the Adam optimizer. The detail of SpecAugment will be presented in Section 5.3.

5.2. Speech-Transformer Model

This is the first time that the Speech-Transformer (sTransformer) has been applied to Vietnamese speech recognition tasks. The Speech-Transformer architecture [19] was inspired by the original Transformer model [16], which has become foundational in natural language processing tasks, including machine translation and language understanding. Unlike traditional sequential models relying on recurrent neural networks (RNNs) or convolutional neural networks (CNNs), Speech-Transformer dispenses with recurrence and convolutions. Note that recurrent networks have limitations in parallelization due to their sequential nature, while convolutional networks are designed for grid-like data such as images. Speech-Transformer targeted to the ASR tasks is a modified version of the Transformer model. Therefore, its architecture is based solely on attention mechanisms which allow the model to focus on different parts of the input sequence when making predictions, providing a more flexible and context-aware approach compared to fixed-size convolutional filters or sequential recurrence. As a result, Speech-Transformer offers advantages in terms of computational efficiency, parallelization, and the ability to capture dependencies across long ranges in sequential data. It has become a cornerstone in the development of SOTA models for ASR tasks.

In this work, we deployed the Speech-Transformer model consisting of a three-layer convolutional neural network (CNN) followed by a Transformer [25] on the LSVSC. The convolutional layers were used to capture local patterns in the spectrogram, and striding was applied to handle the length difference between the input and output sequences. For the three-layer CNN, the following parameters were chosen: CNN input shape = (8, 10, 80), CNN kernel size = (5, 5, 1), CNN stride = (2, 2, 1), CNN output shape = (64, 64, 64). We setup the Transformer with the following parameters: number of encoders = 12, number of decoders = 4, ctc weight for decoder = 0.4, ctc weight for training = 0.3, Transformer dropout = 0.1, regular multihead attention (MHA) with number of heads = 4, d-model = 144, sampling rate = 16 kHz, number of FFT-points = 400, hop length = 160. The model used Noam annealing as learning rate scheduler with and initial learning rate of 0.001 and the number of warmup steps being 25,000. The Adam optimizer was also chosen. To further improve the performance of Speech-Transformer, SpecAugment (cf. Section 5.3) and joint CTC-attention, joint CTC decoding loss (cf. Section 5.4) were integrated. Table 5 summarizes the setup of the main hyper-parameters for the LAS and Speech-Transformer models.

Table 5. Summary of hyperparameter setup for LAS and Speech-Transformer models.

Models	LAS	Speech-Transformer
No. of encoder layers	4	12
No. of decoder layers	2	4
Sampling rate	16 kHz	16 kHz
Window size	400 samples	400 samples
Window stride	160 samples	160 samples
Feature type	log-Mel spectrogram	log-Mel spectrogram
No. of Mel frequencies	80	80
Tokenizer type	word	word
Dropout	0.3	0.1
Learning rate scheduler	Linear warmup (7105 warm-up steps)	Noam annealing (25,000 warm-up steps)
Initial learning rate	0.0003	0.001
Optimizer type	Adam	Adam
No. of epochs	50	100

5.3. SpecAugment

To improve the WER and CER of the LAS and Transformer models on our collected Vietnamese speech corpus, we integrated the fixed SpecAugment technique [26] to the input spectrogram, i.e., a fixed number of time masks and a fixed size of the time mask were selected regardless the length of the input utterance. In our experience, time warping did not contribute to the performance improvement, so we only used time masking, frequency masking and skip time warping. SpecAugment helped to increase the diversity of the dataset as well as the size of the dataset, contributing to solving the overfitting problem that often occurs in deep neural networks.

Although the fixed SpecAugment can significantly improve the system performance of the considered end-to-end ASR models, it may not be adequate for ASR tasks on large-scale corpora that may have a large variance in the length of the input utterances. Therefore, the performance can be further enhanced by using the adaptive SpecAugment technique [27], where the number and size of time masks are adjusted based on the length of the input sequence. Adaptive time masking can be implemented in various ways to adapt to the length of the input spectrogram τ. Here are two different ways:

- Adaptive number of time masks: Adjusting the number of time masks (N_{t-mask}) based on the length of the input spectrogram τ. For example, a rule such as assigning more time masks for longer spectrograms and fewer time masks for shorter ones has been applied. By this way, the masking strategy adapts to the temporal characteristics of the input.

$$N_{t-mask} = \lfloor \rho_N . \tau \rfloor, \tag{1}$$

where ρ_N and τ denote the multiplicity ratio and the time dimension of the input spectrogram, respectively.

- Adaptive size of time masks: Varying the size of the time masks based on the length of the input spectrogram. Longer spectrograms might have larger time masks, while shorter ones might have smaller masks. This adaptive resizing allows the model to selectively focus on different temporal scales depending on the input length.

$$T = \lfloor \rho_S . \tau \rfloor, \tag{2}$$

where ρ_S, T denote the size ratio and the time mask parameter, respectively.

For adaptive time masking, a constant C is used as a upper bound of the number of time masks. Hence, the number of time masks is computed as

$$N_{t-mask} = \lfloor C, \rho_N . \tau \rfloor. \tag{3}$$

5.4. CTC Joint Decoding

Additionally, we utilized a joint CTC-attention model [15] and a joint decoding loss [28] to address the alignment problem, resulting in an enhancement in the system performance. The joint loss function, a combination of connectionist temporal classification loss ($\mathcal{L}_{ctc}$) and sequence-to-sequence loss ($\mathcal{L}_{seq2seq}$), was employed for the training stage, i.e.,

$$\mathcal{L}_{training} = \alpha * \mathcal{L}_{ctc} + (1 - \alpha) * \mathcal{L}_{seq2seq}, \tag{4}$$

where α was a tunable parameter $0 \leq \alpha \leq 1$. In our experiment, $\alpha = 0.3$ was chosen.

At the inference stage, we applied a joint decoding loss that combined the CTC and attention-based sequence probabilities [28]. The joint decoding loss is defined as follows:

$$\mathcal{L}_{decoding} = \beta * \mathcal{L}_{ctc} + (1 - \beta) * \mathcal{L}_{seq2seq}, \tag{5}$$

where β is a tunable parameter $0 \leq \beta \leq 1$. In our experiment, firstly, $\beta = 0.4$ was chosen; then, we evaluated the Speech-Transformer model with different β to find the value of β giving the highest CER and WER (cf. Section 6.3).

6. Simulation Results

In this section, the effectiveness of the LSVSC is evaluated using the LAS and Speech-Transformer (sTransformer) models described in Section 5. The LSVSC was randomly separated into three subsets for training, validation and test, with a ratio of 80:10:10, respectively. There was no overlap among the respective training, validation and test sets. We adopted a log-Mel spectrogram as the input feature for all considered models; the number of Mel frequencies was 80. The most suitable number of epochs for training each model was chosen based on multiple experiments to find out the number that provides the best WER. Therefore, LAS and its variants were trained for 50 epochs, whereas Speech-Transformer and its variants were trained for 100 epochs. In practical terms, we realized that ASR models using the word tokenizer obtained better performance than those using the character tokenizer for Vietnamese speech. The reason is that Vietnamese is a monosyllabic language. Therefore, we selected a word tokenizer instead of a character tokenizer for those models.

6.1. Scenario 1: Evaluate the Mentioned Models for the ASR Tasks on the LSVSC

In this scenario, we compare the performance of LAS, LAS with fixed SpecAugment (LAS + fixedSA), Speech-Transformer and Speech-Transformer with fixed SpecAugment (sTransformer + fixedSA) in terms of CER and WER. For models using fixed SpecAugment, two time masks with time mask factor T = 40 and two frequency masks with frequency mask factor F = 20 were chosen. Table 6 shows the CERs and WERs of the mentioned ASR models for validation (val) and test. We observe that the LAS model obtains good CERs and WERs for both validation and test sets. The CERs of LAS + fixedSA is reduced by 0.5% and 0.7% for the validation and test, respectively, compared to the LAS. The WERs of LAS + fixedSA are also improved approximately 1–1.2%. As we expected, the Transformer model outperforms the LAS models. For the validation and test sets, the Transformer + fixedSA achieves CERs of 4.26% and 4.17%, and WERs of 7.37% and 7.24%, respectively. By integrating fixed SpecAugment, the performance of both the LAS and Transformer models is significantly improved on the LSVSC.

Table 6. Comparison of CER and WER of mentioned ASR models. The best values are highlighted in bold.

Models	CER Val (%)	WER Val (%)	CER Test (%)	WER Test (%)
LAS	5.74	9.74	5.79	9.73
LAS + fixedSA	5.24	8.80	5.05	8.53
sTransformer	4.77	8.22	4.68	8.01
sTransformer + fixedSA	**4.26**	**7.37**	**4.17**	**7.24**

6.2. Scenario 2: Evaluate the sTransformer Model with Different Hyperparameters of adaptSA

6.2.1. Different Hyperparameters for Adaptive Time Mask

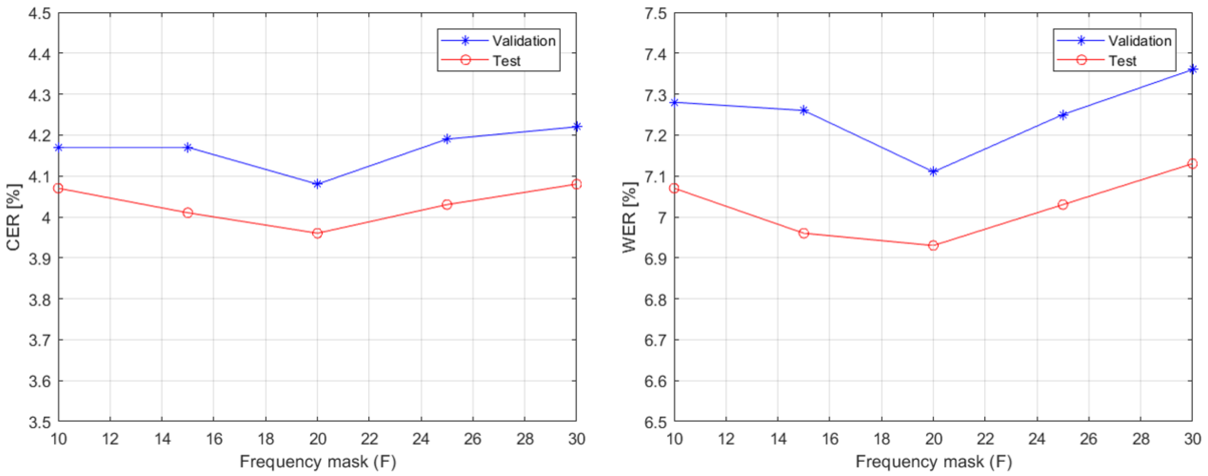
We changed parameters such as ρ_N, ρ_S and C to obtain different adaptive time masks, and then applied them to the input spectrogram of the sTransformer + adaptSA model. The frequency mask factor F = 20 was kept unchanged. For simplicity, we chose $\rho_N = \rho_S = \rho$. Table 7 shows the CERs and WERs of the sTransformer + adaptSA model with different sets of time mask parameters. We see that the performance of this model can be enhanced when we choose suitable parameters for time mask. With the parameter set $[\rho, C] = [0.04, 5]$ the Transformer + adaptSA model achieves the lowest CERs and WERs, specifically CER = 4.08% and WER = 7.11% for the validation set and CER = 3.96% and WER = 6.93% for the test set.

Table 7. Evaluation of sTransformer with different parameter sets of adaptive time mask. The best values are highlighted in bold.

sTransformer + adaptSA	CER Val (%)	WER Val (%)	CER Test (%)	WER Test (%)
$\rho = 0.02, C = 10$	4.22	7.36	4.11	7.17
$\rho = 0.03, C = 10$	4.18	7.28	4.06	7.06
$\rho = 0.04, C = 10$	4.10	7.16	3.97	6.96
$\rho = 0.02, C = 5$	4.24	7.33	4.09	7.14
$\rho = 0.03, C = 5$	4.20	7.32	4.09	7.09
$\rho = 0.04, C = 5$	**4.08**	**7.11**	**3.96**	**6.93**
$\rho = 0.05, C = 5$	4.11	7.16	3.97	6.95

6.2.2. Different Hyperparameters for Frequency Mask

We select the best parameters of time masks, i.e., $\rho = \rho_N = \rho_S = 0.04, C = 5$, and keep those parameters unchanged. We change the parameter of frequency mask (F) from 10 to 30 with a step of 5 to find out the value of F corresponding to the best performance of the sTransformer + adaptSA model. Figure 5 illustrates CERs and WERs of the sTransformer + adaptSA model with different values of frequency masks. It can be seen that with F = 20, this model obtains the best CERs and WERs for both validation and test sets.

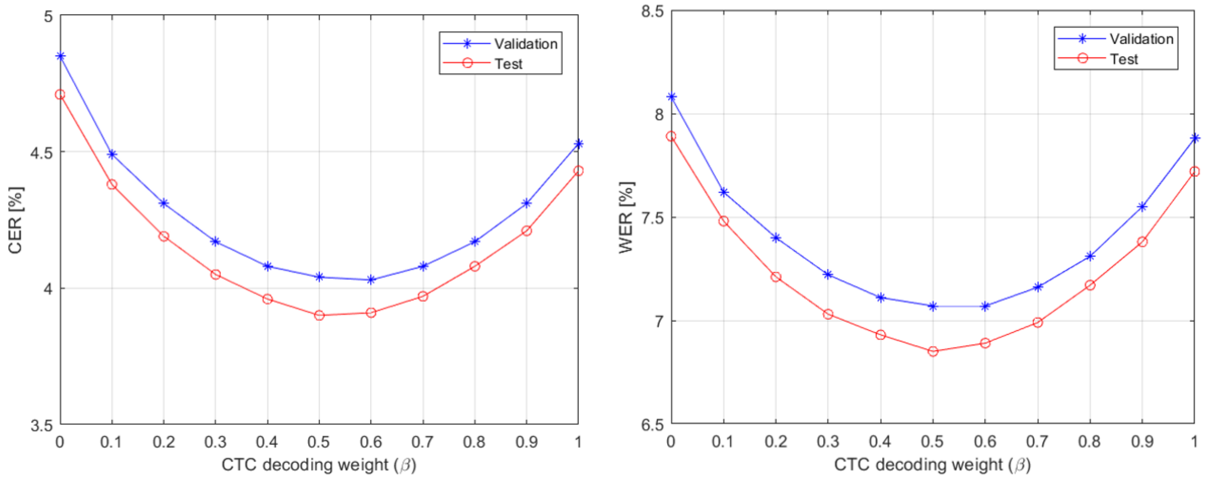


(a) (b)

Figure 5. Performance of the sTransformer + adaptSA model with different values of frequency mask (F): (**a**) CER; (**b**) WER.

6.3. Scenario 3: Evaluate sTransformer + adaptSA Model with Different Ctc Decoding Weight

In this scenario, we keep the best set of parameters for frequency mask and adaptive time mask, i.e., $\rho = \rho_N = \rho_S = 0.04, C = 5, F = 20$ for the sTransformer + adaptSA model. Then we evaluate this model with various values of ctc decoding weight (β), where the ctc decoding weight (as in Equation (5)) is chosen in the range of $[0, 1]$ with step 0.1. The aim is to find the adequate value of β in order to achieve the best performance. Figure 6 shows the performance of the sTransformer + adaptSA model with different values of β. It can be observed that the sTransformer + adaptSA model achieves the lowest CER and WER for both validation and test sets when $\beta = 0.5$. Particularly, with $\beta = 0.5$, we achieve CER = 4.04%, WER = 7.07% for validation and CER = 3.9%, WER = 6.85% for testing.



(a) (b)

Figure 6. Performance of the sTransformer + adaptSA model with different values of ctc decoding weight (β): (**a**) CER; (**b**) WER.

6.4. Scenario 4: Evaluate the sTransformer + adaptSA Model According to Different Categories of LSVSC

6.4.1. Different Regional Dialects

In this part, we evaluate the best sTransformer + adaptSA model ($\rho = \rho_N = \rho_S = 0.04, C = 5, F = 20, \beta = 0.5$) on recognizing Vietnamese speech according to main regional dialects such as the Northern, Central and Southern dialects. Table 8 shows the performance of the sTransformer + adaptSA model for each dialect. The recognition of Northern dialect obtains the lowest CERs and WERs, i.e., CER (val) = 3.76%, WER (val) = 6.62% and CER (test) = 3.65%, WER (test) = 6.47%. Recognizing the Central dialect is more difficult than the Northern dialect but still easier than Southern dialect. The reason is that the prevalence of the Northern dialect as the standard dialect and its widespread usage in the media contribute to its dominance in the collected corpus, comprising approximately 88% of the total data. Although the data of Southern dialect is double the data of the Central dialect, the CERs and WERs in recognizing the Southern dialect are worse than the Central dialect. This is understandable, because the Central dialect is quite similar to the Northern dialect, while the Southern dialect is much different. In spite of difficulty in recognizing the Southern dialect, the sTransformer + adaptSA model still obtains quite good CERs and WERs, for example, CER (test) = 6.15%, WER (test) = 10.57% for the Southern dialect. This experiment confirms that our new corpus has good linguistic diversity across regions. The results correctly reflect the difficulties in recognizing different dialects in Vietnamese speech.

Table 8. Evaluation of sTransformer + adaptSA model according to regional dialects. The best values are highlighted in bold.

Regional Dialects	CER Val (%)	WER Val (%)	CER Test (%)	WER Test (%)
Northern	**3.76**	**6.62**	**3.65**	**6.47**
Central	5.85	9.23	5.43	8.98
Southern	5.96	10.71	6.15	10.57

6.4.2. Different Topics

In this part, we evaluate the best sTransformer + adaptSA model ($\rho = \rho_N = \rho_S = 0.04, C = 5, F = 20, \beta = 0.5$) on recognizing the different types of conversational topics. For simplicity, we categorize the topics of reading (including storytelling and audio books) and news into simple domain of topics, and the topics of healthcare, sports and tourism into the complex domain of topics. The topics in the simple domain are quite popular, and often use common words, but the topics in the complex domain are less common, and may contain specialized terminologies. The simple and the complex domains account for approximately 88% and 5.9% of the our collected LSVSC, respectively. Table 9 demonstrates the CERs and WERs of the sTransformer + adaptSA model according to the simple and complex domains of topics. As expected, the samples in the simple domain of topics are recognized more easily than those in the complex domain. For the simple domain, the sTransformer + adaptSA model performs very well with CER (val) = 3.8%, WER (val) = 6.65% and CER (test) = 3.65%, WER (test) = 6.47%. For the complex domain, the WER (val) and WER (test) are approximately 2.6% and 3.4% higher than those in the simple domain, respectively. This experiment confirms that our new corpus has good diversity in topics. The results reflect the difficulties well in recognizing different domains of topics.

Table 9. Evaluation of sTransformer + adaptSA model according to two main domains of topics.

Domains of Topics	CER Val (%)	WER Val (%)	CER Test (%)	WER Test (%)
Simple domain	3.80	6.65	3.66	6.46
Complex domain	5.32	9.27	5.73	9.85

7. Discussion

In this study, we introduce a new large-scale Vietnamese speech corpus (LSVSC) with 100.5 h of clean speech and transcription. This corpus is open-access. This helps to increase the number of large-scale Vietnamese speech corpora and would become a good dataset for research purposes. Note that Vietnamese is an under-resourced language. The corpus was constructed by crawling untranscribed Vietnamese audio from various sources in the Internet. Then, the raw collected data were preprocessed and transcribed manually with a double check to ensure the high quality of both speech and transcription. In fact, the analyzed results showed that the LSVSC had good diversity of gender and regional dialects and covers multiple types of realistic topics. Moreover, we applied some SOTA end-to-end ASR models like LAS, Speech-Transformer and their variants to evaluate the new corpus. It is the first time those SOTA end-to-end models are used for Vietnamese ASR tasks. Simulation results showed that the mentioned models worked well on the LSVSC in multiple scenarios. The results reflected well the difficulties in recognizing different domains of topics and regional dialects. This implies that the new corpus is a good dataset for Vietnamese ASR tasks.

Actually, we used on-the-fly SpecAugment, i.e., the SpecAugment was applied to the input spectrogram of the model when it was read. The augmented data were not saved after the training stage. Therefore, the size of the training dataset was unchanged. However, the time for training may increase a bit due to the masking process in the SpecAugment. We used NVIDIA GeForce RTX 4090 GPU to compute the average time for training per epoch that the sTransformer with and without SpecAugment require. The results show that the sTransformer with and without SpecAugment need approximate 206.65 s and 200.28 s per epoch, respectively. It means the computational complexity of the sTransformer with SpecAugment is about 3.18% higher than the one without SpecAugment. The time for inference does not change because the SpecAugment is only applied to the input spectrogram during the training stage.

To evaluate the performance of the model when encountering speech features or dialects not present in the corpus, we conducted an experiment to test the sTransformer + adaptSA model using the VIVOS dataset. We obtained CER = 9.25% and WER = 18.44%. The decrease in accuracy was understandable for the unknown data.

In future work, we intend to build a language model (LM) based on the LSVSC. We also want to extend this work by integrating an LM, then apply this model to various applications such as Vietnamese speech-controlled robots, Vietnamese chatbots, Vietnamese voice control in cars, etc.

The source code for implementing simulations in this manuscript can be found at https://github.com/daisankalaeral/VietnameseASR/tree/main, accessed on 1 March 2024.

Author Contributions: L.T.T.T.'s contributions include conceptualisation, methodology, investigation, software, writing—original draft preparation, project administration and funding acquisition. H.-G.K.'s contributions include methodology and writing—review and editing. H.M.L.'s contributions include software and validation. S.V.P.'s contributions include formal analysis, data curation and resources. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Naver corporation, Republic of Korea and the Posts and Telecommunications Institute of Technology (PTIT), Hanoi, Vietnam, grant number 03-PTIT-NAVER-2022.

Data Availability Statement: Our new corpus (LSVSC) can be found at the following link: https://drive.google.com/drive/folders/1tiPKaIOC7bt6isv5qFqf61O_2jFK8ZOI?usp=sharing, accessed on 1 March 2024.

Conflicts of Interest: Han-Gyu Kim is an employee of Naver Cloud, Republic of Korea. 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. Le, V.B.; Tran, D.D.; Castelli, E.; Besacier, L.; Serignat, J.F. Spoken and Written Language Resources for Vietnamese. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, 26 May–1 June 2004; Volume 4, pp. 599–602.
2. Vu, N.T.; Schultz, T. Vietnamese large vocabulary continuous speech recognition. In Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Merano, Italy, 13–17 December 2009; pp. 333–338.

3. Luong, H.T.; Vu, H.Q. A non-expert Kaldi recipe for Vietnamese speech recognition system. In Proceedings of the 3rd International Conference on Experimental and Computational Mechanics in Engineering, Banda Aceh, Indonesia, 11–12 October 2016; pp. 51–55.
4. Nga, C.H.; Li, C.T.; Li, Y.H.; Wang, J.C. A Survey of Vietnamese Automatic Speech Recognition. In Proceedings of the 2021 9th International Conference on Orange Technology (ICOT), Tainan, Taiwan, 16–17 December 2021; pp. 1–4.
5. VLSP 2018—Automatic Speech Recognition, VLSP Home Page. Available online: https://www.vlsp.org.vn/vlsp2018/eval/asr (accessed on 8 March 2018).
6. Nguyen, Q.B.; Mai, V.T.; Le, Q.T.; Dam, B.Q.; Do, V.H. Development of a Vietnamese large vocabulary continuous speech recognition system under noisy conditions. In Proceedings of the 9th International Symposium on Information and Communication Technology (SoICT 2018), Danang, Vietnam, 6–7 December 2018; pp. 222–226.
7. Quoc Truong, D.; Phuong, P.N.; Tung, T.H.; Mai, L.C. Development of high-performance and large-scale vietnamese automatic speech recognition systems. *J. Comput. Sci. Cybern.* **2018**, *34*, 335–348. [CrossRef]
8. Le, V.B.; Tran, D.D.; Besacier, L.; Castelli, E.; Serignat, J.F. First steps in building a large vocabulary continuous speech recognition system for Vietnamese. In *Proceedings of the RIVF 2005*; CiteSeerX: Can Tho, Vietnam, 2005.
9. Nguyen, Q.B.; Dam, B.Q.; Le, M.H. Development of a Vietnamese speech recognition system for Viettel call center. In Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Bali, Indonesia, 8–10 November 2017; pp. 1–5.
10. Nguyen, Q.M.; Nguyen, T.B.; Pham, N.P.; Nguyen, T.L. VAIS ASR: Building a conversational speech recognition system using language model combination. *arXiv* **2019**, arXiv1910.05603.
11. Huy Nguyen, V. An end-to-end model for Vietnamese speech recognition. In Proceedings of the 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (IEEE-RIVF ICCA), Danang, Vietnam, 25–28 November 2019; pp. 1–6.
12. Chorowski, J.; Bahdanau, D.; Cho, K.; Bengio, Y. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv* **2014**, arXiv:1412.1602.
13. 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 (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4960–4964.
14. Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakel, P.; Bengio, Y. End-to-end attention-based large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4945–4949.
15. Kim, S.; Hori, T.; Watanabe, S. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4835–4839.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
17. Chiu, C.C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E. State-of-the-art speech recognition with sequence-to-sequence models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778.
18. Chan, W.; Jaitly, N.; Le, Q.V.; Vinyals, O. Listen, attend and spell. *arXiv* **2015**, arXiv:1508.01211.
19. Dong, L.; Xu, S.; Xu, B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5884–5888.
20. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
21. Graves, A.; Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1764–1772.
22. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G. Deep speech 2: End-to-end speech recognition in english and mandarin. In Proceedings of the International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; pp. 173–182.
23. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv1409.0473.
24. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 577–585.
25. Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J. SpeechBrain: A general-purpose speech toolkit. *arXiv* **2021**, arXiv2106.04624.
26. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019.

27. Park, D.S.; Zhang, Y.; Chiu, C.C.; Chen, Y.; Li, B.; Chan, W.; Le, Q.V.; Wu, Y. Specaugment on large scale datasets. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6879–6883.
28. Watanabe, S.; Hori, T.; Kim, S.; Hershey, J.R.; Hayashi, T. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1240–1253. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.