Pertanika J. Sci. & Technol. 32 (4): 1545 - 1563 (2024)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Low Resource Malay Dialect Automatic Speech Recognition Modeling Using Transfer Learning from a Standard Malay Model

Tien-Ping Tan[1]*, Lei Qin[1], Sarah Flora Samson Juan[2] and Jasmina Yen Min Khaw[3]

[1]School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia
[3]Faculty of Information and Communication Technology, Universiti Tun Abdul Rahman, 31900 Kampar, Perak, Malaysia

ABSTRACT

Approaches to automatic speech recognition have transited from Hidden Markov Model (HMM)-based ASR to deep neural networks. The advantages of deep neural network approaches are that they can be developed quickly and perform better given large language resources. Nevertheless, dialect speech recognition is still challenging due to the limited resources. Transfer learning approaches have been proposed to improve speech recognition for low resources. In the first approach, the model is pre-trained on a large and diverse labeled dataset to learn the acoustic and language patterns from the speech signal. Then, the model parameters are updated with a new dataset, and the pre-trained model is fine-tuned on a low-resource language dataset. The fine-tuning process is usually completed by freezing the pre-trained layers and training the remaining layers of the model on the low-resource language corpus. Another approach is to use a pre-trained model to capture the compact and meaningful features as input to the encoder. Pre-training in this approach usually involves using unsupervised learning methods to train models on a corpus of large amounts of unmarked data. It enables the model to learn the general patterns and relationships between the input speech signals. This paper proposes a training recipe using transfer learning and Standard Malay models to improve automatic speech recognition for Kelantan and Sarawak Malay dialects.

Keywords: Automatic speech recognition, Malay dialects, Malay language, transfer learning

ARTICLE INFO

Article history:
Received: 25 June 2023
Accepted: 18 December 2023
Published: 16 July 2024

DOI: https://doi.org/10.47836/pjst.32.4.06

E-mail addresses:
tienping@usm.my (Tien-Ping Tan)
qinlei@student.usm.my (Lei Qin)
sjsflora@unimas.my (Sarah Samson Juan)
khawym@utar.edu.my (Jasmina Yen Min Khaw)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

The performance of automatic speech recognition (ASR) has advanced rapidly in the past few decades, where the methods have transited from Hidden Markov Model (HMM)-based ASR to hybrid HMM/DNN (deep neural network) models, and now the focus is on end-to-end (E2E) deep neural networks. E2E deep neural networks have several advantages over other models. In general, the E2E deep neural networks can be developed faster compared to other models. For instance, the benchmark HMM-based models in Kaldi ASR (Povey et al., 2011) were developed by Johns Hopkins University and other institutions over an extended period since 2009. In contrast, a deep neural network model can be developed by a person in a few days. In addition, the performance of neural network models is also better than that of HMM when sufficient data is available for training (Watanabe et al., 2017). In an E2E deep neural network model, the joint modeling allows the entire system to be optimized to minimize the overall error, whereas in HMM-based ASR errors from one module can propagate through the system and accumulate, potentially degrading performance. Despite this, developing an ASR that performs well in low-resource languages is still very challenging. One of the challenges in low-resource ASR is dialect speech recognition.

Dialects occur as a result of culture, customs, and geography. With urbanization and social development, several dialects have become endangered or are spoken by only a small community. It means that most dialect corpora, if available, are small. In addition, it is difficult to transcribe dialect speech because no standard writing system exists for most dialects. Thus, dialect speakers who want to communicate in text may use different spelling rules. Hence, dialect speech recognition is a crucial and challenging problem in low-resource ASR.

One of the languages with many dialects is Malay. Malay belongs to the Austronesian family and is designated as the official language of Malaysia, Indonesia, Singapore, and Brunei. The Malay languages spoken in these countries may differ in pronunciation and vocabulary and are considered dialects. However, most Malay dialects do not have a written form. The formal Malay language recognized in Malaysia is Standard Malay, which originates from the Johor-Riau dialect (Asmah, 1991). The Johor-Riau dialect gained prominence due to the influence and importance of the empire during the 19th century. Malay dialects in Malaysia can be categorized based on their geographical distribution (Colins, 1989). The Malay dialects in Peninsular Malaysia are classified into seven groups: (1) the North-Western group, which includes Kedah, Perlis, Penang, and North Perak dialects; (2) the North-Eastern group, which is the Kelantan dialect; (3) the Eastern group, which is the Terengganu dialect, (4) the Southern group, which comprises Johor, Melaka, Selangor, and Perak (Southern), (5) the Negeri Sembilan group, (6) the Pahang dialect as a separate group, and (7) the Perak dialect, which covers the area of Central Perak.

Table 1 shows some text samples of Kelantan and Sarawak Malay compared to Standard Malay. There is no formal orthography for Malay dialects. The native speakers will write the dialect words based on how they are pronounced with reference from Standard Malay (Khaw et al., 2024). From these examples, we can see that the grammar of the dialect is similar. However, there are some insertions, deletions, and substitutions of letters in the dialect Malay words compared to the Standard Malay words. In addition, there are also unique vocabularies in Malay dialects that do not exist in Standard Malay.

Table 1
Example of sentences in Kelantan dialect and Sarawak dialects and their translation in Standard Malay

Malay Dialects	**Standard Malay**
kalu keno tange kito keno kulit mesti la gata (Kelantan Malay)	kalau kena tangan kita kena kulit mestilah gatal
cucuk pertama.. nunggu kitak lambat gilak. malas nak berbini (Sarawak Malay)	cucu pertama.. tunggu kamu lambat sangat. malas nak beristeri

There are many similarities between the Malay dialect and Standard Malay. In this study, we investigate using transfer learning in an end-to-end deep neural network to improve the performance of dialect Malay automatic speech recognition, specifically in Kelantan Malay and Sarawak Malay dialects.

Malay Automatic Speech Recognition

There are a few studies on Malay automatic speech recognition. However, most of the works used HMM as their models. For example, Tan et al. (2008) trained a large vocabulary HMM/GMM ASR for read speech using Sphinx 3 ASR and obtained a WER of 14.6%. Chong et al. (2012) collected a Malay broadcast news and trained an HMM/GMM ASR using the Kaldi toolkit and obtained a WER of 17.1%. Juan et al. (2012) analyzed the speech recognition of Malay, Chinese, and Indian speakers and concluded that native Malay speakers have a lower WER compared to non-native speakers. Rahman et al. (2014) studied the Malay ASR for children. Their proposed approach achieves a WER of 24%.

Dialect Automatic Speech Recognition

In general, automatic speech recognition models the acoustics, pronunciation, and language or word sequence given speech utterances and their respective transcription. In an HMM-based ASR, they are trained or built separately in different models: language, pronunciation, and acoustic (Koehn et al., 2007). A language model such as n-gram captures the linguistic context of a speech by modeling the relationship between words or sub-word tokens. A pronunciation dictionary normally models the relationship between the words and their pronunciations on phones, while an acoustic model contains the phones and their respective

acoustics features. However, in an E2E deep neural network ASR, such as the encoder-decoder model (Hori et al., 2017), the acoustics, pronunciation, and language representation are learned in a single deep neural network.

Several approaches can be applied to improve dialect ASR performance. First, collecting and augmenting dialect speech data can improve the result tremendously. It uses augmentation techniques such as pitch shifting, noise addition, and speed perturbation to create additional training examples (Renduchintala, 2018; Aitoulghazi et al., 2022). The approach is able to increase the size of the speech corpus more than one-fold. Second, studies show that acoustic modeling involving unsupervised learning of multilingual speech and transfer learning can improve ASR performance (Baevski et al., 2020). Third, modeling dialect words and their respective phones through grapheme-to-phoneme (G2P) can also improve accuracy. Fourth, including dialect-specific vocabulary, phrases, and text to model a separate language model for an E2E ASR can also be useful. This adaptation can involve incorporating dialect-specific language resources or adapting existing models using dialectal text data. Ali (2020) proposed to use a deep neural network consisting of a convolutional neural network (CNN), recurrent neural network, and a 4-gram language model for Arabic dialect speech recognition. A CTC beam search decoder guided by an n-gram language model was used for decoding. Next, we will focus on works that applied acoustic modeling approaches to improve the E2E ASR.

Acoustic Modeling in Dialect Automatic Speech Recognition

A typical strategy to improve the performance of dialect ASR is to integrate dialect information into the model. Li et al. (2018) proposed using an encoder-decoder neural network to train an ASR model for seven English dialects: America, India, Britain, South Africa, Australia, Nigeria, Ghana, and Kenya. The authors proposed appending a tag that contains the dialect information to the transcription. It allows the system to perform automatic speech recognition and dialect classification at the same time. They also proposed using cluster adaptive training for their model. The size of the speech corpus in the study is very large, with about 40 thousand hours of noisy training data consisting of 35 million utterances. Compared with the dialect-independent models, the proposed model improves the word error rate (WER) by 1%–3%. Grace et al. (2018) proposed a similar approach, including the dialect information in the feature vector instead. They showed that the proposed model outperformed dialect-specific models.

Jain et al. (2018) proposed approaching the problem using dialect embedding inspired by x-vectors in speaker recognition (Snyder et al., 2018). They extracted the dialect embedding from a standalone time-delay neural network (TDNN) dialect classifier. The dialect embedding was used to augment the speech feature vectors consisting of MFCC and i-vectors. The authors trained a TDNN that jointly performs speech recognition and

dialect classification. The WER also improved in the range of 1%–3%. The approach of integrating dialect information is interesting. Nevertheless, this approach is suitable for training models with large speech resources.

Transfer learning has been proven to be an effective method to improve the performance of low-resource language ASR tasks because it allows models to learn from larger and more diverse data sets and transfer the knowledge to low-resource tasks. In E2E ASR, transfer learning can improve the performance of low-resource language tasks by leveraging the learned knowledge from high-resource language tasks. It is achieved by using a pre-trained model to improve the performance of the target model.

There are two approaches to implementing transfer learning. In the first approach, the model is pre-trained on a large and diverse labeled dataset to learn the acoustic and language patterns from the speech signal. Then, the model parameters are updated with a new dataset, and the pre-trained model is fine-tuned on a low-resource language dataset. The fine-tuning process is usually completed by freezing the pre-trained layers and training the remaining layers of the model on the low-resource language corpus.

Yan et al. (2018) used a Time Delay Neural Network with Long Short-Term Memory Projection (TDNN-LSTMP) for Tibetan dialects speech recognition. The speech corpus consists of Tibetan dialects U-Tsang and Amdo, with 62 hours and 52 hours of speech, respectively. The authors proposed using the Mandarin TDNN-LSTMP, trained using 1700 hours of speech, as the pre-trained model for Tibetan speech recognition. The experiments show that Mandarin to U-Tsang can achieve a remarkable performance, and the U-Tsang to Amdo is also effective. The approach obtains a relative improvement in WER, around 1%–4%.

Hou et al. (2020) transferred the large-scale E2E model trained using multilingual speech corpora from 42 languages as the pre-trained model to 14 low-resource languages. They used an encoder-decoder based on a transformer with hybrid CTC/attention for the ASR and language identification tasks. The pre-trained model contains speech data for around 1 to 7 hours. Their model achieved significantly superior results to the non-pre-trained baseline on the language-specific low-resource ASR task. The average WER for the 14 languages decreased from 83.4% to 60% for language-specific models.

Another approach is to use a pre-trained model to capture the compact and meaningful features as input to the encoder. Pre-training in this approach usually involves using unsupervised learning methods to train models on a corpus of large amounts of unmarked data. It enables the model to learn the general patterns and relationships between the input speech signals. One of the pre-trained models that received wide attention is the Wav2Vec2 (Baevski et al., 2020). The Wav2vec2 is the enhanced version of the Wav2Vec model that was trained using a self-supervised learning approach using a large dataset of 53,000 hours of unlabeled speech data from 53 languages. Once the model is trained on a large dataset, the representations learned by the acoustic encoder can be used as a starting point

for downstream speech recognition tasks. The idea is similar to fine-tuning large language models such as BERT for different natural language processing tasks. Baevski et al. (2020) trained a Wav2Vec2 model using LibriVox, and the pre-trained model was then fine-tuned with only 10 minutes of labeled data. Testing the trained model with Librispeech obtained a WER of 5.2% on the clean speech. The result is very promising, showing that ASR can be trained and performed well using a few minutes of speech data.

The raw speech is fed into a convolutional neural network (CNN) called the acoustic encoder to train the Wav2Vec2 model. The acoustic encoder then converts the raw audio waveform into a sequence of intermediate representations called acoustic features. The acoustic features are subsampled to reduce the temporal resolution. Context window masking encourages the model to learn representations that capture context. In this process, random subsequences of the subsampled acoustic features are masked, and the model is trained to predict the masked-out portions. The masked acoustic features are then passed through a stack of Transformer layers. The Transformer model learns to capture the relationships between different parts of the masked acoustic features and generate contextualized representations. The contextualized representations from the Transformer layers are used to calculate a contrastive loss. This loss encourages similar samples to have representations that are close together in the embedding space while pushing dissimilar samples apart. The model is trained using stochastic gradient descent (SGD). Figure 1 shows the Wav2Vec2 model.

Yi et al. (2021) applied Wav2Vec2 in speech recognition and obtained 20%–50% relative improvements in WER for low-resource languages. They used the CALLHOME telephone conversation speech corpus, which consists of six languages, namely Mandarin, English, Japanese, Arabic, German, and Spanish, with approximately 15 hours of labeled speech for fine-tuning.

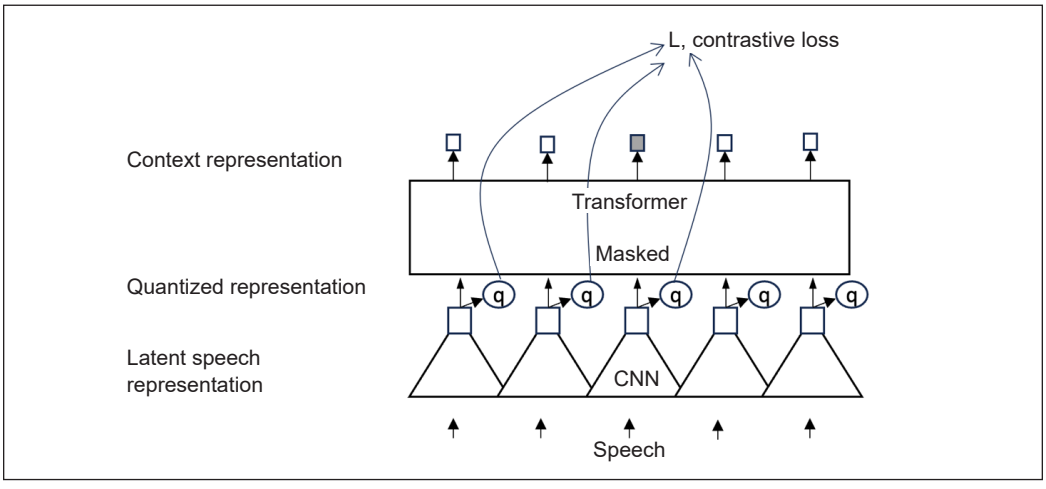

Figure 1. Wav2Vec2 (Baevski et al., 2020)

MATERIALS AND METHODS

We first discuss the resources used in the study; then, we briefly analyze dialects in terms of the phonemes and the writing used in the dialects. Next, we propose the E2E ASR model used and the training steps to improve the E2E Malay dialect ASR model.

Malay Dialect Speech Corpus

This study aims to investigate and propose an automatic speech recognition modeling for Malay dialects. We used the Malay dialect conversation speech corpus by Khaw et al. (2024) for training and testing. The speech corpus consists of dialogs in Kelantan Malay and Sarawak Malay. The corpus consists of many dialog conversations, each consisting of two Malay speakers who discuss a topic of interest for ten minutes in Malay dialect. The conversation was captured through the headset and recorded using the CoolEdit software, with the recording being set at 16 kHz/16 bits per sample. The speech was later transcribed to text by native speakers in the spoken dialect using Praat and translated into Standard Malay. Figure 2 shows a speech utterance transcribed in Kelantan Malay and Standard Malay. Table 2 describes the Malay dialect speech corpus.

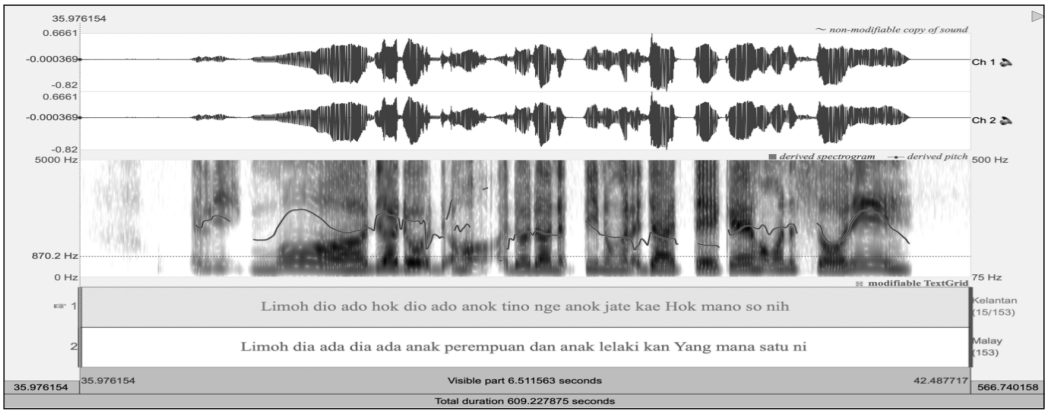

Figure 2. Praat was used to transcribe the conversation (Boersma, 2001)

Table 2
Malay dialect conversation speech corpus

Description	Recorded Speech Conversation	
	Kelantan	**Sarawak**
Age	21–24	31
Gender	9 female, 1 male	1 female, 1 male
Total Duration	1 hour 40 mins	1 hour and 20 mins
Number of Dialogs (2 speakers/dialog)	10	8
Training	1 hour 20 mins, 8 speakers	1 hour, 2 speakers
Testing	20 mins, 2 speakers	20 mins, 2 speakers

Malay Speech Corpus

Besides the Malay dialect speech corpus, we used a standard Malay speech corpus, MASS speech corpus, in our study. The MASS speech corpus (Tan et al., 2009) is a Standard Malay read speech corpus that consists of 199 speakers and about 140 hours of speech. Refer to Table 3 for information about the Standard Malay speech corpus speakers. About 120 hours of speech is used for training, while 20 hours is used for testing. The speakers consist of Malay, Chinese, Indian, and others.

Table 3
MASS speech corpus (Tan et al., 2009)

Speakers		Number	Hours	Age (mean, min, max)	Gender
Training	Malay	66	39	25.5, 10, 51	38F, 27M
	Chinese	98	73.5	27.3, 16, 38	65F, 34M
	Indian	8	5.5	24.4, 23, 28	1F, 7M
	Others	3	2	29.3, 28, 32	1F, 2M
Testing	Malay	8	6	28.8, 10, 48	4F, 4M
	Chinese	14	12	27.1, 20, 33	7F, 7M
	Indian	2	1.5	24.5, 23, 26	1F, 1M

Linguistics

In terms of linguistics, Sarawak Malay and Kelantan Malay have similar phoneme sets to Standard Malay. Specifically, Sarawak Malay has all the phonemes in Standard Malay except /e/. Kelantan Malay also has all the phonemes in Standard Malay, and in addition, it has double consonants, which are not available in Standard Malay. Table 4 below shows the Standard Malay, Sarawak Malay, and Kelantan Malay phoneme sets. Nevertheless, the vocabulary used in Standard Malay and Malay dialects are different. The dialect speakers will insert, replace, or delete one or more letters in a Standard Malay word due to the difference in the pronunciation or phoneme in the dialect word. For instance, in Table 1, Kelantan dialect speakers insert the letter "k" into the word "*cucu.*" In addition, Kelantan dialect speakers delete the letter "l" from the word "*gatal*." At the same time, Sarawak dialect speakers delete the second letter "*a*" from the word "*kalau*." Kelantan dialect speakers may also substitute Standard Malay words that end

Table 4
Comparing the Standard Malay and Malay dialect phonemes (Khaw, 2017)

	Phoneme
Standard Malay	p, b, t, d, k, g, ʔ, s, x, h, f, v, z, ʃ, tʃ, ʤ, l, r, m, n, ŋ, ɲ, w, j, a, ə, e, i, o, u, ai, au, oi
Sarawak Malay	p, b, t, d, k, g, ʔ, s, x, h, f, v, z, ʃ, tʃ, ʤ, l, r, m, n, ŋ, ɲ, w, j, a, ə, i, o, u, ai, au, oi
Kelantan Malay	p, b, t, d, k, g, ʔ, s, x, h, f, v, z, ʃ, tʃ, ʤ, l, r, m, n, ŋ, ɲ, w, j, a, ə, e, i, o, u, ai, au, oi, pp, bb, tt, dd, kk, gg, ss, cc, jj, ll, mm, nn, ww

with “a” with “o.” In contrast, the Sarawak Malay speaker pronounces the Standard Malay word “*tunggu*” as “nunggu.”

On the other hand, if we analyze the distribution of letters in Sarawak Malay, Kelantan Malay, and Standard Malay, we can see that the letters in the dialect and Standard Malay have a positive correlation. Specifically, from Figure 3, we can see that Kelantan Malay has a lower percentage of “a” but a higher percentage of “k,” “o,” and “t” compared to Standard Malay. We can observe, for instance, in the Standard Malay words that end with the letter “a” or the last syllable that has a vowel “a,” the words in Kelantan Malay will substitute the letter “a” with “o.” Thus, words like “*nama*” and “*semak*” are written as “*namo*” and “*semok*”. Contrary to Sarawak Malay, the dialect has a lower percentage of “e,” “i,” “n,” “r,” and “u” but a higher percentage of “k” compared to Standard Malay. For Sarawak Malay, it can be observed that many words end with the letter “k” compared to Standard Malay.

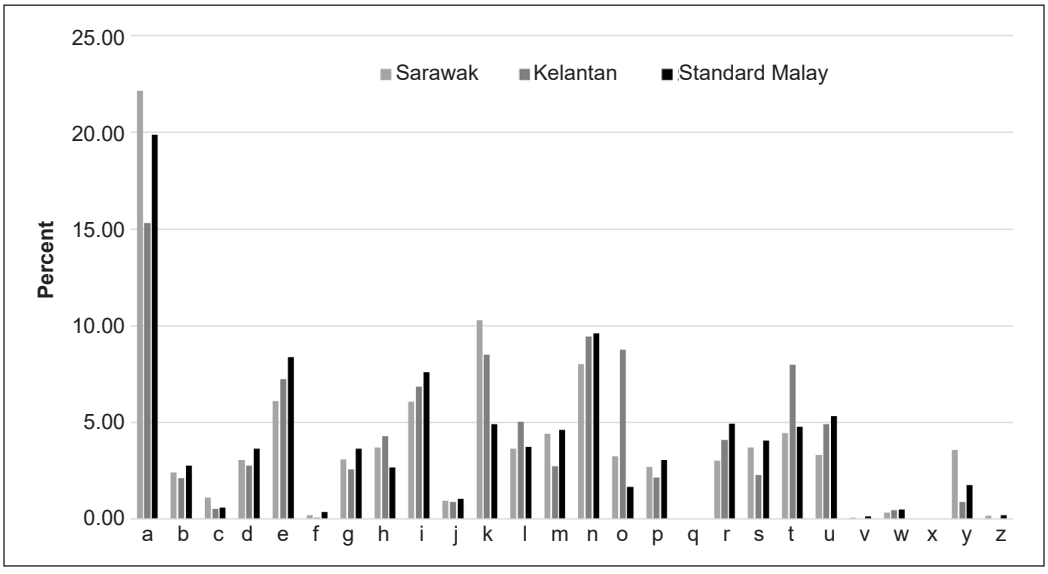

Figure 3. Distribution of letters in Malay Dialects and Standard Malay

Since there are many overlaps in the phoneme set, a positive correlation of letters between Malay dialect and Standard Malay, and abundant Standard Malay resources, we are interested in investigating using Standard Malay speech corpus to improve the dialect Malay speech corpus.

E2E ASR Model

We propose an E2E deep neural network model that consists of a Wav2Vec2, an encoder model, and a classifier using CTC loss, as shown in Figure 4, for our automatic speech recognition. Wav2Vec2 is a pre-trained model that was self-supervised and trained using

a large amount of speech corpora. As for the encoder, we propose an encoder that consists of three deep neural network blocks, where each block consists of a linear layer, batch normalization layer, dropout layer, and leaky ReLU layer, as opposed to a more complex encoder such as a transformer encoder or encoder-decoder architecture for modeling Malay dialect because there are less than two hours of speech data available for training. The output layer is a linear layer of size 36 that will decode speech features to letters using Connectionist Temporal Classification (CTC) loss.

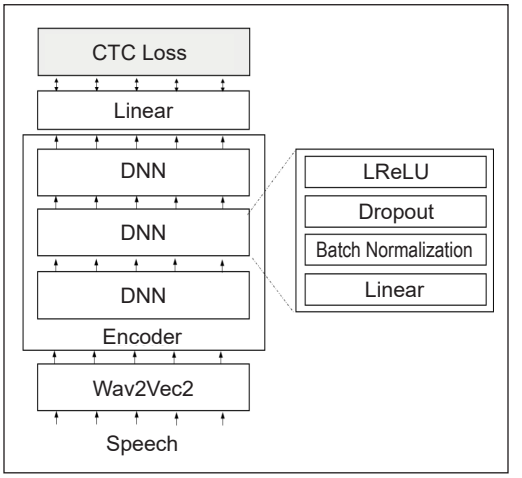

Figure 4. Wav2Vec2 and the encoder model were used

Since the amount of training data available is very small, we performed speech perturbation at the rate of 0.95% and 1.05% to the speech utterances to triple the amount of speech data available. We propose to train the pre-trained Wav2Vec2 and encoder model using Standard Malay speech data. Figure 5 shows our proposed modeling step. First, we used the training set of the MASS corpus, which consists of about 120 hours of Standard Malay read speech corpus to train the pretrained Wav2Vec2 (Wav2Vec2-XLSR-53) and

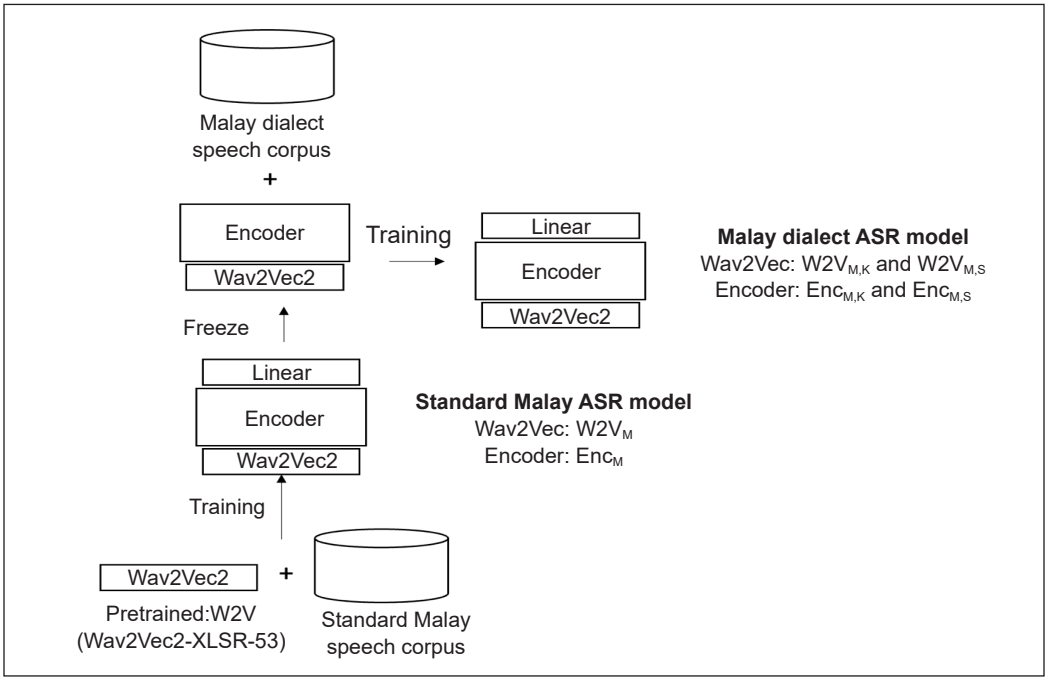

Figure 5. Proposed steps for Malay dialect ASR modeling

DNN model. It produces the trained Wav2Vec2 ($W2V_A$) and encoder (Enc_A) models. Besides, we also investigate the effect of using only the Standard Malay speech from native Malay speakers in the training. Thus, the subset of the MASS corpus that contains 39 hours of Standard Malay from native Malay speakers was used to train the pre-trained Wav2Vec2 and DNN model, which resulted in the $W2V_M$ and Enc_M models. The weights of the Wav2Vec2 model and the encoder model produced were frozen. We then retrained the frozen models (either the Wav2Vec2 model or both the Wav2Vec2 model and encoder model) with dialect speech to produce Malay dialect ASR models. In Figure 5, we can use the Kelantan Malay dialect speech to train the pre-trained $W2V_M$ model and the untrained encoder model to obtain the $W2V_{M,K}$ and Enc_K models. Refer to the naming convention in Table 5 used for training the models.

Table 5
The naming convention is used to train the different dialect models

Standard Malay speech	Dialect	Wav2Vec2	Encoder
Native and non-native Malay	-	$W2V_A$	Enc_A
Native Malay	-	$W2V_M$	Enc_M
-	Kelantan	$W2V_K$	Enc_K
-	Sarawak	$W2V_S$	Enc_S
Native and non-native Malay	Kelantan	$W2V_{A,K}$	$Enc_{A,K}$
Native Malay	Kelantan	$W2V_{M,K}$	$Enc_{M,K}$
Native and non-native Malay	Sarawak	$W2V_{A,S}$	$Enc_{A,S}$
Native Malay	Sarawak	$W2V_{M,S}$	$Enc_{M,S}$

RESULTS AND DISCUSSION

The experiments used the Malay dialect conversation speech corpus (Khaw et al., 2024). Refer to Table 2 for information about the Malay dialect conversation speech corpus. The Malay dialect conversation speech recording and transcription were segmented using ffmpeg and script. For the training and testing and Kelantan Malay automatic speech recognition, eight dialogs from eight speakers (consisting of 1 hour 20 mins) were used for training (where two speakers took part in a dialog). In contrast, two separate dialogs from two speakers were used (consisting of 20 mins) for testing. There was no overlap in the speakers in the training and testing dataset. Since the amount of data available for training and testing is small, we used a cross-validation strategy to evaluate our model, where we rotated the training and testing data, as shown in Table 6. On the other hand, the training of Sarawak Malay was carried out using six dialogs (consisting of 1 hour), while two dialogs containing 20 mins were used for testing. Speakers overlap in training and testing for the Sarawak Malay dataset because we only have two speakers. The same cross-validation strategy was used to evaluate our model (Table 6).

Table 6
Cross-validation strategy used to evaluate the model. The training and testing data are split as follows

Dialect	Training set	Test set	Training Speakers	Testing Speakers
Kelantan Malay	Dialog 3, 4, 5, 6, 7, 8, 9 & 10	SET 1: Dialog 1 & 2	C, D, E, F, G, H, I, J	A, B
	Dialog 1, 2, 5, 6, 7, 8, 9 & 10	SET 2: Dialog 3 & 4	A, B, E, F, G, H, I, J	C, D
	Dialog 1, 2, 3, 4, 7, 8, 9 & 10	SET 3: Dialog 5 & 6	A, B, C, D, G, H, I, J	E, F,
	Dialog 1, 2, 3, 4, 5, 6, 9 & 10	SET 4: Dialog 7 & 8	A, B, C, D, E, F, I, J	G, H
	Dialog 1, 2, 3, 4, 5, 6, 7 & 8	SET 5: Dialog 9 & 10	A, B, C, D, E, F, G, H	I, J
Sarawak Malay	Dialog 3, 4, 5, 6, 7 & 8	SET 1: Dialog 1 & 2	K, L	K, L
	Dialog 1, 2, 5, 6, 7 & 8	SET 2: Dialog 3 & 4	K, L	K, L
	Dialog 1, 2, 3, 4, 7 & 8	SET 3: Dialog 5 & 6	K, L	K, L
	Dialog 1, 2, 3, 4, 5 & 6	SET 4: Dialog 7 & 8	K, L	K, L

We were interested to know if there is any difference between Standard Malay and Standard Malay from only native Malay for the task. Thus, we used the MASS speech corpus (Tan et al., 2009), a Malay read speech corpus, for training the Standard Malay ASR models. The speechbrain toolkit was used in the experiments (Ravanelli et al., 2021). We trained two ASR models: a Standard Malay model using all the training sets of the MASS corpus ($MASS_A$) and another Standard Malay model using only the Malay native speaker speech ($MASS_M$). The $MASS_A$ described in Figure 1 was trained using about 120 hours of speech consisting of 191 speakers from the training set. Note that the pre-trained Wav2Vec2 model used in this training was Wav2Vec2-XLSR-53 from Huggingface. The result is a Standard Malay model: Wav2Vec2 model ($W2V_A$) and encoder model (Enc_A). To verify the effectiveness of the model proposed in Figure 1, we tested the model using about 20 hours of Standard Malay speech that consists of 22 speakers from the test set.

The word error rate (WER) was 16%. The result is very good, considering that we do not use any language model in the decoding, and the model used is simpler compared to an encoder-decoder model. We trained another Standard Malay model using about 39 hours of speech from 66 native Malay speakers. We trained the pre-trained Wav2Vec2-XLSR-53 model using the native Standard Malay speech in this case. The result is a Standard Malay model: Wav2Vec2 model ($W2V_M$) and encoder model (Enc_M). The Wav2Vec2 produces about 50 acoustic feature vectors for one second of speech. The size of the linear layer in the DNN was set at 1024, while the output linear layer was set at 36. The batch size is 6, and Adam optimizer was used. The maximum number of epochs was set at 30. The learning rate is set at 10^{-3}. Greedy decoding was used to predict the best path for each utterance.

As described earlier, a cross-validation strategy was used in the experiments, where we rotate the speakers for training and testing. There were five test sets in the Kelantan Malay

experiment and four in the Sarawak Malay experiment, each consisting of two dialogs totaling 20 minutes. For the first baseline model, we tested the dialect models $W2V_A$+Enc_A and $W2V_M$+Enc_M on the dialect speech. Both models produced more than 90% WERs. The results show that the Standard Malay models cannot recognize the dialect words. Next, we trained the dialect ASR models depicted in Figure 5 using Kelantan dialect and Sarawak dialect speech, respectively, and the models produced are denoted as $W2V_K$+Enc_K and $W2V_S$+Enc_S, respectively. The initial pre-trained Wav2Vec2 model used in both cases was Wav2Vec2-XLSR-53. The baseline Kelantan Malay model and Sarawak Malay model have an average word error rate (WER) of 68.2% and 69.4%, respectively. Refer to Table 7 for the Kelantan Malay result and Table 8 for the Sarawak Malay result.

The results show that using a small amount of Malay dialect speech in training produces a better ASR model than Malay read speech. Nevertheless, the high WER was due to the very low resources used in the training, and the speech was spontaneous conversation. Nevertheless, the results were very good compared to the models trained using Standard Malay speech, where the WERs were more than 90%. We conducted three experiments to investigate the modeling steps proposed in Figure 5. In the first experiment, we train the pre-trained Standard Malay Wav2Vec model ($W2V_A$) that we described earlier using the training set of Kelantan Malay speech and Sarawak Malay speech to produce the Kelantan Malay model ($W2V_{A,K}$) and Sarawak Malay model ($W2V_{A,S}$), respectively. The encoders were trained from scratch. Thus, the encoders produced for Kelantan Malay and Sarawak Malay were Enc_K and Enc_S, respectively. We tested the models using the Malay dialect speech test set using the same cross-validation strategy.

There is an improvement in the WER for Kelantan Malay and Sarawak Malay speech compared to the baseline, where the average WER was reduced to 65.6% and 62.1%, respectively. In the second experiment, we retrained the pre-trained Wav2Vec2 ($W2V_A$) and pre-trained encoder (Enc_A, which were trained using Standard Malay speech). We obtained the Kelantan Malay model ($W2V_{A,K}$+$Enc_{A,K}$) and the Sarawak Malay model ($W2V_{A,S}$+$Enc_{A,S}$), respectively. We examined the models using the Malay dialect test sets using the cross-validation strategy, and the results show that the WER of both the Kelantan and Sarawak models slightly deteriorated compared to using only the Wav2Vec2 model. The result shows that pretraining Wav2Vec2 with Standard Malay speech can improve the WER of Malay dialect ASR, but pretraining the encoder with Standard Malay speech is not useful.

We repeated the previous experiment using the Wav2Vec2 pre-trained with Standard Malay speech from native speakers, which were $W2V_M$. Specifically, in experiment 3, we trained the pre-trained $W2V_M$ using Kelantan Malay and Sarawak Malay speech and obtained $W2V_{M,K}$ and $W2V_{M,S}$, respectively. Subsequently, we tested them using their respective test set. Interestingly, both models show lower WER compared to those trained using pre-trained $W2V_A$ and those trained using native and non-native Malay speech. The

average WER for Kelantan Malay ASR drops more than 4%, while the average WER for Sarawak ASR improves by more than 7% compared to the baseline. Nevertheless, the improvement in the WER of Sarawak Malay ASR from using $W2V_{M,S}$ is lower compared to the improvement in the WER of Kelantan Malay ASR from $W2V_{M,K}$. We did not evaluate the Enc_M since, in the previous experiment, combining the encoder showed no improvement in the WER.

First, the experiments show that training the pre-trained Wav2Vec2 using Malay speech is beneficial, as more discriminative features can be produced during feature extraction, which improves the WER. Furthermore, the results show that the pre-trained Wav2Vec2 that was trained with native Standard Malay speech, $W2V_M$, then used for training dialect models $W2V_{M,K}$ and $W2V_{M,K}$ give lower WERs compared to $W2V_{A,K}$ and $W2V_{A,K}$, respectively. The non-native Malay speech may not be useful in adapting Wav2Vec2 to the Malay dialect ASR. On the other hand, since the writing of the dialect speakers is not the same, using the encoder model that is fine-tuned using Malay during training does not give a significant advantage, which can be seen in a drop in WER. The fourth observation is that the improvement in WER for the Malay dialect with a similar phoneme set to Standard Malay shows a higher improvement in WER when the native Malay Wav2Vec2 was used. Tables 7 and 8 show the complete results obtained in the experiments.

Table 7
WER of the Kelantan Malay testing sets

	Kelantan Malay				
Test set	**Baseline: $W2V_A$+Enc_A**	**Baseline: $W2V_K$+Enc_K**	**$W2V_{A,K}$+ Enc_K**	**$W2V_{A,K}$+ $Enc_{A,K}$**	**$W2V_{M,K}$+ Enc_K**
SET 1	97.0%	66.4%	58.0%	58.5%	**57.2%**
SET 2	97.7%	68.4%	65.9%	66.1%	**64.3%**
SET 3	96.4%	64.0%	63.9%	64.9%	**60.5%**
SET 4	98.2%	76.3%	75.0%	75.1%	**74.1%**
SET 5	97.0%	65.9%	65.4%	65.0%	**63.9%**
Average	97.3%	68.2%	65.6%	65.9%	**64.0%**

Table 8
WER of the Sarawak Malay testing sets

	Sarawak Malay				
Test set	**Baseline: $W2V_A$+Enc_A**	**Baseline: $W2V_S$+Enc_S**	**$W2V_{A,S}$+ Enc_S**	**$W2V_{A,S}$+ $Enc_{A,S}$**	**$W2V_{M,S}$+ Enc_S**
SET 1	95.2%	69.1%	65.4%	66.4%	**64.0%**
SET 2	93.5%	65.9%	58.5%	59.1%	**58.4%**
SET 3	93.4%	75.1%	**63.4%**	66.4%	63.6%
SET 4	95.1%	67.6%	61.3%	63.2%	**60.8%**
Average	94.3%	69.4%	62.1%	63.8%	**61.7%**

We further evaluate if the improvement is significant between the baselines and the proposed models ($W2V_{M,K}$+Enc_K or $W2V_{M,S}$+Enc_S). We applied the T-Test known as the Matched Pairs Sentence-Segment Word Error (MAPSSWE) test to evaluate whether two ASR performances in WER are significantly different. Specifically, the null hypothesis is that the two systems have no performance difference. We used the Speech Recognition Scoring Toolkit (SCTK) version 2.4.12 (https://www.nist.gov/itl/iad/mig/tools) from the National Institute of Standards and Technologies (NIST), United States, to perform the tests. We performed a separate test for the Kelantan and Sarawak decoding/hypotheses. Since we experimented using cross-validation, the decoding from the respective model has to be merged and compared against the reference. For the Kelantan dialect models, the minimum p that the test finds a significant difference between $W2V_K$+Enc_K and $W2V_{M,K}$+Enc_K is p=0.004 (99.6% confidence level), which is lower than the standard p=0.05 (or higher than the 95% confidence level). On the other hand, for Sarawak dialect models, the minimum p that the test finds a significant difference between $W2V_S$+Enc_S and $W2V_{M,S}$+Enc_S is p=0.001 (99.9% confidence level). Thus, both tests show that the results from the proposed models are statistically significant compared to the baseline results.

In addition, we select an utterance from Kelantan Malay and an utterance from Sarawak Malay for further analysis and discussion. We use word alignment for analysis as it is the commonly used approach in ASR. We conducted word alignment of the hypothesis and reference. The Kelantan Malay utterances were decoded using $W2V_M$+Enc_M and $W2V_{M,K}$+Enc_K. The following is the alignment of the decoding output (`Kelantan_001_005_002_063.wav`) that was produced using $W2V_M$+Enc_M (hypothesis) compared to the reference:

```
============================================================
Kelantan_001_005_002_063.wav, %WER 100.00 [ 6 / 6, 0 ins, 3
del, 3 sub ]
Ref: nok ;       gi      ; banyok ; banyok ;  tuh  ;  gok
      S  ;       S       ;   S    ;   D    ;   D   ;   D
Hyp: nur ; gibayebaannya ; tuguh  ; <eps>  ; <eps> ; <eps>
============================================================
```

The word alignment shows that the WER is 100%, which means none of the words produced is correct. However, we can see many characters are correct if we analyze the character alignment. The CER for the $W2V_A$+Enc_A is about 50% for the Kelantan Malay dialect and Sarawak Malay dialect using $W2V_A$+Enc_A. The result generally means that one in two characters are wrongly decoded, which is quite good compared to the word error rate. One of the reasons is due to the unseen phone in the Malay dialect. However, the average WER for the Kelantan Malay dialect and Sarawak Malay dialect of more than 90% show that the model has a problem with word segmentation. It is similar to human

auditory perception of an unknown language or dialect, where we have difficulty identifying words in the utterance even though we know all the sounds. When the model was trained with dialect speech, the same utterance was decoded by $W2V_{M,K}$+Enc_K to produce the following output:

```
============================================================
Kelantan_001_005_002_063.wav, %WER 16.67 [ 1 / 6, 0 ins, 0
del, 1 sub ]
Ref: nok ; gi ; banyok ; banyok ; tuh ; gok
      =  ; =  ;   =    ;   =    ;  S  ;  =
Hyp: nok ; gi ; banyok ; banyok ;  tu ; gok
============================================================
```

The WER of the utterance reduces to 16.67%, which means the dialect speech helps the model recognize unique phones in the dialect and word segmentation. Next, we analyze the alignment of the decoding output (Sarawak-03-96.wav) that was produced using $W2V_M$+Enc_M (hypothesis) compared to the reference:

```
============================================================
Sarawak-03-96.wav, %WER 125.00 [ 5 / 4, 1 ins, 0 del, 4 sub ]
Ref: umo  ; baru ; brapa  ; bulan ; <eps>
      S   ;  S   ;   S    ;   S   ;   I
Hyp: emua ;  wa  ; operwo ;  ladn ; hinis
============================================================
```

The WER of the utterance is at 125%, which is very high. In the reference, we can see that some Sarawak Malay words are also Standard Malay words, such as "*baru*" and "*bulan*," but the $W2V_M$+Enc_M model still has difficulty identifying these words. One of the reasons is that the dialect of Malay speech we are decoding is conversation speech, which is more spontaneous compared to read speech. Many ASR studies have shown that conversation speech has a higher WER compared to read speech. Second, the Standard Malay speech corpus is a read speech corpus. Thus, a mismatch in the training and testing speaking style may affect the decoding. After the model is trained with Kelantan Malay speech, the $W2V_{M,S}$+Enc_S model produces the following output. The WER of the utterance reduces to 50%, and the number of correctly decoded characters also increases.

```
============================================================
sarawak-03-96.wav, %WER 50.00 [ 2 / 4, 0 ins, 0 del, 2 sub ]
umo  ; baru ; brapa ; bulan
 S   ;  S   ;   =   ;   =
rumo ; maok ; brapa ; bulan
============================================================
```

CONCLUSION

In this study, we propose training steps that use the Wav2Vec2 that was fine-tuned using Standard Malay for subsequent modeling with Malay dialect speech for automatic speech recognition. The Wav2Vec2 that was fine-tuned using Standard Malay speech from native Malay speakers showed improved WER/CER for Kelantan and Sarawak Malay dialects. In contrast, the Standard Malay speech from all the speakers, including non-native speakers, only produces lower improvement in WER/CER in the Sarawak Malay speech recognition task. At the same time, there is no improvement in WER/CER in the Kelantan Malay speech recognition task.

The improvement in WER and CER is encouraging from Wav2Vec2; nevertheless, it is higher than the studies conducted by Yi et al. (2021) and Baevski et al. (2020). One possible reason may be that the amount of dialect speech used for training was smaller compared to Yi et al. (2021). Besides, we do not get a single-digit WER compared to Baevski et al. (2020) because the languages the authors tested were already modeled in the Wav2Vec2. On the other hand, the Malay dialects that were evaluated were not modeled in the Wav2Vec2. Nevertheless, Wav2Vec2 was trained in some Malay speech. The results show that the languages used to train the Wav2Vec2 affect the accuracy of the ASR. In addition, non-native Standard Malay speech does not seem useful in the Malay dialect ASR transfer learning.

ACKNOWLEDGMENT

Oracle supported this research with a Research Grant. Award Number: CPQ-2746169, and Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101007666.

REFERENCES

Aitoulghazi, O., Jaafari, A., & Mourhir, A. (2022). DarSpeech: An automatic speech recognition system for the Moroccan dialect. In *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)* (pp. 1-6). IEEE Publishing. https://doi.org/10.1109/ISCV54655.2022.9806105

Ali, A. R. (2020). Multi-dialect Arabic speech recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE Publishing. https://doi.org/10.1109/IJCNN48605.2020.9206658

Asmah, H. O. (1991). *Aspek bahasa dan kajiannya* [Aspects of language and its study]. Dewan Bahasa dan Pustaka.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, Article 2006.11477. https://doi.org/10.48550/arXiv.2006.11477

Boersma, P. (2001). Praat: A system for doing phonetics by computer. *Glot International*, *5*(9) 341-345.

Chong, T. Y., Xiao, X., Xu, H., Tan, T. P., Chau-Khoa, P., Lyu, D. C., Chng, E. S., & Li, H., (2013). The development and analysis of a Malay broadcast news corpus. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)* (pp. 1-5). IEEE Publishing. https://doi.org/10.1109/ICSDA.2013.6709862

Colins, J. T. (1989). Malay dialect research in Malaysia: The issue of perspective. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 235-264.

Grace, M., Bastani, M., & Weinstein. E. (2018). Occam's adaptation: A comparison of interpolation of bases adaptation methods for multi-dialect acoustic modeling with LSTMS. In *2018 IEEE Spoken Language Technology Workshop (SLT),* (pp. 174-181). IEEE Publishing. https://doi.org/10.1109/SLT.2018.8639654

Hori, T., Watanabe, S., Zhang, Y., & Chan, W. (2017). Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. *ArXiv,* Article 1706.02737. https://doi.org/10.48550/arXiv.1706.02737

Hou, W., Dong, Y., Zhuang, B., Yang, L. Shi, J. & Shinozaki, T. (2020). Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. In *Interspeech* (pp. 1037-1041). ISCA Publishing. https://doi.org/10.21437/Interspeech.2020-2164

Jain, A., Upreti, M. & P. Jyothi, P. (2018) Improved accented speech recognition using accent embeddings and multi-task learning. In *Proceedings of Interspeech* (pp. 2454-2458). ISCA Publishing.

Juan, S. S., Besacier, L., & Tan, T. P. (2012). Analysis of Malay speech recognition for different speaker origins. In *2012 International Conference on Asian Language Processing* (pp. 229-232). IEEE Publishing. https://doi.org/10.1109/IALP.2012.23

Khaw, J. Y. M. (2017). *Bootstrapping Kelantan and Sarawak Malay dialect models on text and phonetic analyses in text-to-speech system* [Doctorate Dissertation]. Universiti Sains Malaysia.

Khaw, J. K. M., Tan, T.-P. & Ranaivo-Malancon, B. (2024). Hybrid distance-statistical-based phrase alignment for analyzing parallel texts in standard Malay and Malay dialects. *Malaysian Journal of Computer Science, 37*(1), 89–106. https://doi.org/10.22452/mjcs.vol37no1.5

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions,* (pp. 177-180). Association for Computational Linguistics.

Li, B., Sainath, T. N., Sim, K. C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y. & Rao, K. (2018) Multi-dialect speech recognition with a single sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 4749-4753). IEEE Publishing. https://doi.org/10.1109/ICASSP.2018.8461886

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G. & Vesel, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* (pp. 1-4). IEEE Signal Processing Society.

Rahman, F. D., Mohamed, N., Mustafa, M. B., & Salim, S. S. (2014). Automatic speech recognition system for Malay speaking children. In *2014 Third ICT International Student Project Conference (ICT-ISPC)* (pp. 79-82). IEEE Publishing. https://doi.org/10.1109/ICT-ISPC.2014.6923222

Ravanelli, M., Parcollet, T., Plantinga, P. W., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J., Yeh, S., Fu, S., Liao, C., Rastorgueva, E. N., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., & Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. *ArXiv*, Article 2106.04624. https://doi.org/10.48550/arXiv.2106.04624

Renduchintala, A., Ding, S., Wiesner, M., & Watanabe, S. (2018). Multi-modal data augmentation for end-to-end ASR. *ArXiv*, Article 1803.10299. https://doi.org/10.48550/arXiv.1803.10299

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 5329-5333). IEEE Publishing. https://doi.org/10.1109/ICASSP.2018.8461375

Tan, T. P., Xiao, X., Tang, E. K., Chng, E. S., & Li, H. (2009). MASS: A Malay language LVCSR corpus resource. In *2009 Oriental COCOSDA International Conference on Speech Database and Assessments* (pp. 25-30). IEEE Publishing. https://doi.org/10.1109/ICSDA.2009.5278382

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. In *IEEE Journal of Selected Topics in Signal Processing*, (Vol. 11, No. 8, pp. 1240-1253). IEEE Publishing. https://doi.org/10.1109/JSTSP.2017.2763455

Yan, J., Yu, H., & Li, G. (2018). Tibetan acoustic model research based on TDNN. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 601-604). IEEE Publishing. https://doi.org/10.23919/APSIPA.2018.8659584

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2021). Transfer ability of monolingual Wav2vec2.0 for low-resource speech recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE Publishing. https://doi.org/10.1109/IJCNN52387.2021.9533587