Automatic Multilingual Speech Recognition

Nguyen Tuan Anh , Tran Thi Ngoc Linh , Dang Thi Hien

Abstract - **Automatic Speech Recognition (ASR) for multi-languages is currently attracting more and more attention; however, development is still hampered by the need for language experts. End-to-End ASR simplifies their work by directly predicting the output character based on the acoustic input. This study presents the improvement of LIS-Net model for End-to-End Vietnamese and Chinese ASR system. In this study, an efficient yet accurate end-to-end multilingual multi-speaker ASR model has developed, allowing direct conversion of raw speech audio signals into text of multiple languages. This study proposes a new method of coding labels specifically for multiple languages by pagination labels by language. The results of this study are significantly improved compared to that of baseline models.**

I. INTRDUCTIONS

Automatic Speech Recognition (ASR) systems that can transcribe speech in multiple languages, known as multilingual models [1]. End-to-End Automatic Speech Recognition for multiple languages is one of the most fascinating areas, which has been attracting a lot of attention lately. Although a data hungry, the accuracy of monolingual ASR model has reached par with humans on a number of tasks. [2],[3]. With many languages missing or few training resources, the results are still very limited, however, it is gaining a lot of interest in developing high-performance ASR systems [4], [1]. This suggests that both monolingual conventional systems and monolingual E2E models have Word Error Rate (WER) higher than end-to-end multilingual model in Large-Scale Multilingual dataset. Moreover, a model for multiple languages will significantly reduce infrastructure compared to each model's language. The basic principles for building a successful multilingual model, which have been published today include shared hidden layers [5], stacked bottleneck features [5]–[8], multitask learning [9] and knowledge distillation [5].

The current development of multilingual systems for countries in the Asia-Pacific region has not received adequate attention. The accuracy is very limited because of the peculiarities of languages, dialects, languages of ethnic minorities. Lack of data, lack of model strong enough to increase accuracy. Therefore, in this study, the approach is to create models for multiple languages with a combination of rich and low-resourced languages. From there, the internal bottleneck approach extracts sharing features across languages that are used to cross-train languages [6], [10], [11].

Some main contributions of this study as follows:

- Improved the single model for monosyllabic and multi-tone in multi-languages.
- Auto language identification in the model that does not need to use multiple models for the task of language identification and content recognition.
- Language expandable, with the new label coding method, the model can add other languages to the model.
- Expert grammar knowledge free, with end-to-end model training, it is not necessary to understand grammar rules.

II. RELATED WORK

Multi-language speech recognition has been developed for a long time [12]–[14]. However, in recent years, it has become a phenomenon with outstanding development in application, so there are many directions for strong development of investment, such as:

- Large-scale multilingual ASR ([1], [15])
- Low-resource ASR ([16]–[18])
- Multi-model or Multi-Task for multi-language ASR ([19]–[21])
- Code-Switching multilingual ASR ([22]–[27])

Many previous models have many limitations when working with Multilingual ASR ([28]–[31]). Some phrase recognition on multilingual keyword spotting ([32], [33]) for multilingual is also studied. Studies on Acoustic Models (AMs), some models have focused on the study of common phone sets ([28], [29]), some others have designed models with share parameters ([12],[9]). A design noteworthy in models is that some of the lower layers of the Deep Neural Network (DNN) are shared between languages and the output layer is language-specific ([12], [31], [34]). Traditional models often require language-specific Pronunciation Models (PMs) and Language Models (LMs). Therefore, while inference must know speech language identity ([31]). In addition, AMs, PMs and LMs are usually optimized independently so errors can occur during training that are difficult to control [35].

Recently, a lot of research has focused on developing single end-to-end models for multilingual. These models have many advantages such as replacing AMs, PMs, and LMs of n different languages with a single model while continuing to show improved performance over monolingual E2E systems ([4], [35], [36]).

III. BLOCK-BASED RESIDUAL NEURAL NETWORK (BRN)

A. Language Coding and Identity

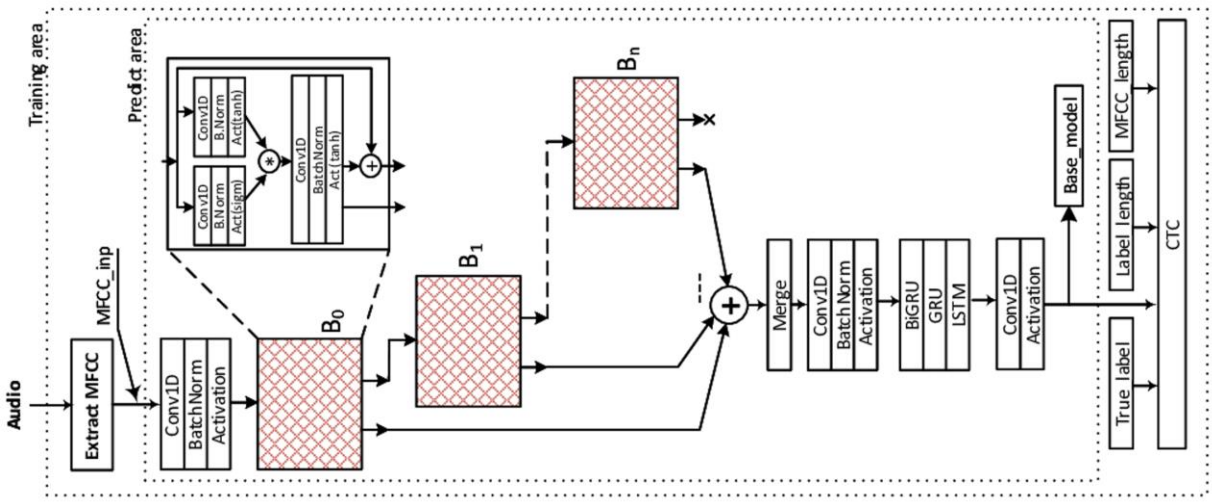

Figure 1. System Network Architecture

For coding, each language will be using word-based. Theoretically, the number of languages that can be coded is arbitrary. To improve the ability to compare results, this study was conducted in two languages, Chinese and Vietnamese. In Chinese, because the language does not use the alphabet, each character is considered a word, a meaningful word can be composed of several words put together, and when compounding words, there can be variations in pronunciation. From there we have two ways to encode, one is coding by each character and the second is coding by each meaningful phrase. To minimize the number of vocabulary words, this study chose word encoding, with Chinese being a single character. Here, all characters in the database are encoded:

$$page_i \rightarrow page_i + n_{w_i} \qquad (1)$$

Where i refers to a language, n_{w_i} refers to the total vocabulary of that language (shown as Table 1). Similar to Vietnamese, the feature of this language is the use of alphabet system and extended characters in Unicode encoding to character encoding, word structure consists of several letters and accented markings for tones. Each word may not have a complete meaning yet. To make sense of what people want to say, some words still need to be put together. Thus, each character in Chinese will be corresponding to the word in Vietnamese. For simplicity of coding, this study chose word coding. From there the Vietnamese data will be coded as formula (1), but located in another page (using new i for each language). Word-based is selected because it was proven that it can balance both OOV and performance issues ([37][38])

Table 1. Languages label coding

Page 0				Page 1				...	Page n			
W0	W1	…	Wn0	W0	W1	…	Wn1	…	W0	W1	…	Wnn

B. *Imbalanced Multilingual Data Processing*

This section is described research methods for balancing data in a multilingual model. Data imbalance is a common phenomenon of speakers in languages around the world. Languages with more speakers will tend to have more data. In ASR systems, an E2E multilingual model will be trained on all components, therefore, the data imbalance is very sensitive. In this section, data balancing is explored, data sampling - ratio pickup method.

Data imbalance often results in a model working better in languages with more data. Suppose we have a multilingual model trained with k languages $L_0, \ldots, L_{k-1}$, where L_i has n_i training samples and $N = \sum_{i=0}^{k} n_i$. Training models are performed for each batch sampled from N samples in the dataset. In each batch, the ratio between languages will be chosen equal to the ratio of the total number of samples between those languages in which L_i has ratio of $s(i) = \frac{n_i}{N}$. his means the model is updated by $s(i)$ times with L_i and $s(j)$ times with $L(j)$ language. The rate of updating the gradient of L_i compared to L_j is $\frac{n_i}{n_j}$.

C. *Baseline Models*

BRN is designed based on the ideas of the cutting-edge models, to understand the architecture of BRN, the theory of these baseline models will be shown below.

ResNet - Residual Network. [39]. In Deep Neural Network, when increasing the number of layers, it will be harder to train. According to K.He in the paper "Deep Residual Learning for Image Recognition" [39], he found that adding more layers to the network would increase training error and harder to train to achieve high accuracy. Deep Residual Convolution significantly increases the number of layers in a network. Input data is processed by layers or a block then add or concatenate with previous layer via shortcut to produce output results. Gradient in ResNet can flow directly from input to output of convolutional layer and/or blocks. On the other hand, due to the base network is convolution, so calculation speed is faster than other structure. He has shown empirical evidence to show that the ResNet network is more easily optimized and can achieve significant increases in depth. In the article “Deep Residual Learning for Image Recognition” [39], they showed that their depth reached 152 layers and won the ILSVRC & COCO 2015 Competitions, has the best results compared to only

Table 2. Statistics of VIVO database

Dataset	Speaker	Male	Female	Utterance	Duration(h)	Unique Syllables
Train	46	22	24	11660	14:55	4617
Test	19	12	7	760	00:45	1692

Table 3. Statistics of THCH-30 dataset

Dataset	Speaker	Male	Female	Age	Utterance	Duration(hour)
Train	30	8	22	20-50	10893	27:23
Test	10	1	9	19-50	24	6:24

stack layer network structures. The structure of ResNet network as Figure 2

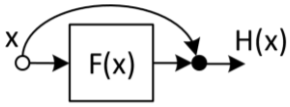

Figure 2. Residual network block

The output of the layers or a block of the network is $H(x)$, If x was fitted to label then simply set the weight to 0, otherwise, fitting $F(x)$, so we have:

$$F(x) = H(x) - x \quad (2)$$

The advantage of this network is to keep features, to avoid vanilla gradient, however, due to regular gradient can only flow along the transmission network and can be flown across or combined with the features in each block by summation, which can degrade information before the end network.

D. Block-Based Residual Neural Network (BRN)

In the ASR pre-processing, we denote X as MFCC input feature, $X = [x_0, x_1, \dots, x_{n-1}]$ where n refer to number of samples in dataset. Each feature comes with a true label, so we denote L as true labels vector, $L = [l_0, l_1, \dots, l_{n-1}]$. A label vector, $l_i, i = 0, .., n-1$, can be vectorized into $l_i = [w_0, w_1, \dots, w_{m-1}]$ where $w_j, j = 0, .., m-1$ refer to each word in the vocabulary, $m = |l_i|$, vectorized method depends on the language to choose character or word based. In this study, only keywords is focused with a given list ω, $\omega = [\omega_1, .., \omega_q]$, with q refer to number of given keywords, all words in labels vector will be treated as garbage (ω_0) if it is not in ω, so we denote $\hat{\omega}$ as KWS labels vector, $\hat{\omega} = \omega_0 \cup \omega$ and $L \in \hat{\omega}$. Noted that ω_0 cannot be treated as Null, None or "blank" (in CTC procedure) because the position of w_j maybe need to observe.

The architecture of the network is defined as shown in the Figure 1. Pre-processing step, denote as "Extract MFCC" block, creating MFCC future of utterances, saving to hard disk if it is not existed, which will help the training process faster in later step; the "Base_model" block refer to predict model. $B_k, k = 0, .., N-1$ is denoted as ResBlock[1] module which plays the role of the core module of the BRN network, N refer to number of $ResBlock * block$. Each of core block B_k is constructed by 1D Convolution, Batch Normalization and Activation layers that is calculated by the formula (3), for simplicity, b is omitted in the equation.

$$H_\delta(x) = \delta(W^T x) \quad (3)$$

where x and H refer to input and output vector of the layers considered. δ or maybe σ represents the activation type of Sigmoid or Tanh. To optimize the parameters, through practical experiments, a dropout layer is added after the Batch Normalization layer. The dropout parameters have been changed when experiments to find the best results. The element-wise multiplication in formula (4 is used in the core block B_k.

$$F(x) = \sigma\left(W\left(H_\sigma(x) * H_\delta(x)\right)\right) \quad (4)$$

Here F is the output vector of residual mapping to be learned with input vector x. To perform shortcut connection for B_k, calculation in (5) is used by element-wise addition.

$$B_k = F(x) + x \quad (5)$$

Here the output, B_k, will be treated as input vector x of B_{k+1}, and $F(x)$ in (6) is aggregated in residual transformation as described as in ResNeXT.

$$R = H_\sigma\left(\sum F(x)\right) \quad (6)$$

where R is extracted feature output of BRN that are shared from multiple layers from input to output of the network. This is an advantage of BRN for the purpose of sharing features in multilingual KWS. Characteristics between different languages will be similar, so sharing features between multiple languages will reduce the number of network parameters. From here, to improve predictive accuracy, several RNN classes will be used.

$$\Psi = g\left(RNNs(R)\right) \quad (7)$$

Where R is the aggregated of blocks in (6). Ψ denote as "Base_model" block as shown in Figure 1, using to predict. $RNNs$ are some Recurrent Neural Network layers, including Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), which can be customized to the appropriate number of layers. The final output, Ψ, is the Probability Feature table, which will be used to calculate the loss with the CTC[40] function to update all parameters.

IV. Experimental Results

A. Datasets

THCH-30 corpus is an open Chinese speech dataset, released by Tsinghua University [41] with a total of 30 hours of speech, recorded in a quiet room. This dataset has the characteristics as in Table 3.

The selection of keywords is done by taking 10 words with the largest frequency of occurrence in the

[1] ResBlock: ResNet block with modified structure

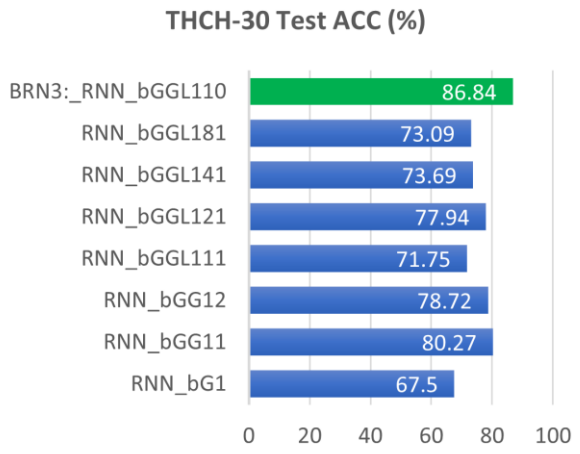

Figure 4. THCH-30 Test ACC (%), higher is better

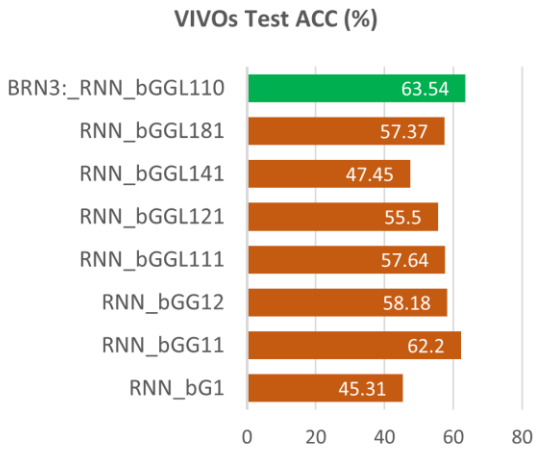

Figure 5. VIVOs Test ACC (%), higher is better

Table 1. Differences model structures configuration

Model Name	Model Architect	Capacity	Total Params
RNN_bG1	BiGRU	64	33K
RNN_bGG11	BiGRU-GRU	64-64	70K
RNN_bGG12	BiGRU-GRU	64-128	132K
RNN_bGGL111	BiGRU-GRU-LSTM	64-64-64	103K
RNN_bGGL121	BiGRU-GRU-LSTM	64-128-64	181K
RNN_bGGL141	BiGRU-GRU-LSTM	64-256-64	411K
RNN_bGGL181	BiGRU-GRU-LSTM	64-512-64	1165K
BRN3:_RNN_bGGL110	BRN-BiGRU-GRU-LSTM	3-64-64-32	292K

entire data set to do the experiments. After analyzing, ten keywords to train was chosen as in list KW:
KW = ['的', '一', '有', '人', '了', '不', '为', '在', '用', '是']

Figure 3. Chinese Keyword list. These words can be read as follows: De, yi, you3, ren4, le, bu2, wei2, zai2, yong2, shi2

VIVOS Corpus is a free Vietnamese Speech dataset [42]. It includes 15 hours of voice recording using for ASR task. Published by AILAB, a computer science laboratory of VNU - Hanoi University of Technology. This is the only one open dataset for ASR in Vietnamese. The characteristics describe as in Table 2. The method of selecting keywords is the same as on THCH-30 dataset.

B. Model Architecture Configurations

The hyper-parameters are optimized by Adam optimizer, the maximum training is about 400 epochs. Batch size varies according to the hardware trained on. To achieve the highest training speed, learning rate is selected from 0.001 then reduced after 5 epochs without improved results. Each utterance input is extracted MFCC with 20-channels. The length of features is padding equal to the longest one. In this case, the maximum padding is 375. To compare the effectiveness of the models, our model and the baseline were experimented with various custom changes.

C. Results and Discussion

To be able to compare with other models, the study [43] was selected as a baseline model because of their experimental results show that LSTM models has been outperformed the feed-forward DNN and performed better compared to cross-entropy loss trained LSTM. Moreover, LSTM, BiLSTM or Multi-layer RNN architecture has been used in many KWS tasks such as ([44]–[50]).

1) Compare block-based residual network with baseline models

The results of this study are shown in Table 1. Specifically, "BRN3: _RNN_bGGL110" is the version of our model that delivers the best results compared to all base model customizations. With RNN, a carefully calibrated customization produces the best results, but it only achieves an ACC score of 80.27% for THCH-30, while for VIVO is 62.2%, still lower than BRN. To better understand the convention of each network option, the options are denoted as follows:
The capacity is the number of hidden cells (with RNN) or block (with BRN). Symbols of models are defined as follows:

- RNN: Recurrent Neural Network
- bG: Bidirectional GRU (Gated recurrent units)
- G: GRU (Gated Recurrent Units)
- L: LSTM (Long Short-Term Memory)
- BRN: Block-based Residual Convolutional Neural Network

Table 1 shows the names of the models and the number of layers established on the network. In each layer, the size of each hidden layer, makes it easy to recreate the model. The “Total params” column tells us about the capacity of the network, in principle, the smaller the better. From this, we can observe that the BRN network is medium in size compared to the variants of the RNN.

From Figure 4 and Figure 5, we can observe that the results of BRN are higher than that of DeepRNN with the same RNN architecture and with the similar training time. For Chinese, because the dataset is about 30h, BRN shows superior results compared to baselines. The highest RNN result was 80.27% and the BRN result was 86.84%, higher than baseline 33.3%.

With Vietnamese, this dataset has nearly 15 hours, too few for a language, so all models are hard to give the best results in practice. However, to compare the BRN with the baseline, once again from Figure 5, the BRN produces a higher result, 63.54% compared to 62.2%. The author hopes that if the amount of data is increased, the models will increase accuracy. It is also a work to be done in the future.

2) Discussion.

As shown in Figure 4 and Figure 5, when using Deep RNNs with varying network depths and using BRN with the varying configurable blocks, it is easy to observe that BRN has better results in most cases. From there, the model can easily use code-switching for multiple languages or code-switching for multi-dialect languages.

V. Conclusions

In this study, we revisited the baseline models architect. Based on these cutting-edge structures, we have introduced the BRN network with design by inheriting those advantage design. The purpose of this network is improving the accuracy when detecting speech keywords in multiple datasets for multiple languages without aligned labels, in this case, the Mandarin and Vietnamese. This Keyword Spotting approach is suitable for real applications. Experiment on THCH-30 and ViVos corpus, both datasets are very modest in quantity. BRN shows that it is outperforms Deep RNNs base models. As future work, we want to expand this work to support simultaneous KWS for multi-language online recognition.

VI. Acknowledgments

In this article, we would like to especially thank to Thai Nguyen University of Technology, Thai Nguyen, Viet nam for supporting us in the experimental process.

VII. References

[1] A. Kannan *et al.*, “Large-scale multilingual speech recognition with a streaming end-to-end model,” *[C] Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 2130–2134, 2019, doi: 10.21437/Interspeech.2019-2858.

[2] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, “The Microsoft 2017 Conversational Speech Recognition System,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, no. August, pp. 5934–5938, 2018, doi: 10.1109/ICASSP.2018.8461870.

[3] W. Xiong *et al.*, “The Microsoft 2016 Conversational Speech Recognition System,” in *Proc. IEEE ICASSP*, 2017, pp. 5255–5259.

[4] S. Tong, P. N. Garner, and H. Bourlard, “Cross-lingual adaptation of a CTC-based multilingual acoustic model,” *[J]Speech Commun.*, vol. 104, pp. 39–46, 2018, doi: 10.1016/j.specom.2018.09.001.

[5] T. Sercu *et al.*, “Network architectures for multilingual speech representation learning,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 5295–5299, doi: 10.1109/ICASSP.2017.7953167.

[6] S. Thomas, S. Ganapathy, and H. Hermansky, “Multilingual MLP features for low-resource LVCSR systems,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4269–4272.

[7] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, “Investigation on cross-and multilingual MLP features under matched and mismatched acoustical conditions,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7349–7353.

[8] J. Cui *et al.*, “Multilingual representations for low resource speech recognition and keyword search,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 259–266.

[9] D. Chen and B. K.-W. Mak, “Multitask learning of deep neural networks for low-resource speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 7, pp. 1172–1183, 2015.

[10] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X. Zhang, “Investigation of Multilingual Deep Neural Networks for Spoken Term Detection,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013.

[11] F. Grézl, M. Karafiát, and K. Vesel\`y, “Adaptation of multilingual stacked bottle-neck neural network structure for new language,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7654–7658.

[12] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7304–7308.

[13] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, “Multilingual deep neural network based acoustic modeling for rapid language adaptation,” in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2014, pp. 7639–7643.

[14] A. Mohan and R. Rose, “Multi-lingual speech recognition with low-rank multi-task deep neural networks,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4994–4998.

[15] C. Liu, Q. Zhang, X. Zhang, K. Singh, Y. Saraf, and G. Zweig, “Multilingual ASR with Massive Data Augmentation,” *arXiv Prepr. arXiv1909.06522*, Sep. 2019.

[16] R. Sahraeian and D. Van Compernolle, “A study of rank-constrained multilingual DNNs for low-resource ASR,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5420–5424.

[17] S. Zhou, S. Xu, and B. Xu, “Multilingual End-to-End Speech Recognition with A Single Transformer on Low-Resource Languages,” *arXiv Prepr. arXiv1806.05059*, pp. 2–6, 2018.

[18] R. Menon, H. Kamper, E. Van Der Westhuizen, J. Quinn, and T. Niesler, “Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 3475–3479, 2019, doi: 10.21437/Interspeech.2019-1665.

[19] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe, “Multi-modal data augmentation for end-to-end ASR,” in *[C] Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, vol. 2018-Septe, pp. 2394–2398, doi: 10.21437/Interspeech.2018-2456.

[20] Z. Tang, L. Li, and D. Wang, “Multi-task recurrent model for true multilingual speech recognition,” *2016 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2016*, 2017,

doi: 10.1109/APSIPA.2016.7820821.

[21] M. A. Di Gangi, M. Negri, and M. Turchi, "One-To-Many Multilingual End-to-end Speech Translation," *arXiv Prepr. arXiv1910.00254*, Oct. 2019.

[22] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, "Towards End-to-End Code-Switching Speech Recognition," *arXiv Prepr. arXiv1810.12620*, 2018.

[23] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the end-to-end solution to Mandarin-English code-switching speech recognition," in *[C] Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, vol. 2019-Septe, pp. 2165–2169, doi: 10.21437/Interspeech.2019-1429.

[24] Y. Khassanov *et al.*, "Constrained output embeddings for end-to-end code-switching speech recognition with only monolingual data," in *[C] Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, vol. 2019-Septe, pp. 2160–2164, doi: 10.21437/Interspeech.2019-1867.

[25] M. A. Menacer, D. Langlois, D. Jouvet, D. Fohr, O. Mella, and K. Smaïli, "Machine Translation on a Parallel Code-Switched Corpus," *[L] Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11489 LNAI, pp. 426–432, 2019, doi: 10.1007/978-3-030-18305-9_40.

[26] E. Yılmaz, A. Biswas, E. Van Der Westhuizen, F. De Wet, and T. Niesler, "Building a unified code-switching ASR system for South African languages," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, pp. 1923–1927, 2018, doi: 10.21437/Interspeech.2018-1966.

[27] C. Shan *et al.*, "Investigating End-to-end Speech Recognition for Mandarin-english Code-switching," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 6056–6060, 2019, doi: 10.1109/ICASSP.2019.8682850.

[28] T. Schultz and A. Waibel, "Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets," in *Proceedings of the 5th European Conference on Speech Communication and Technology, Vol. 1*, 1997, pp. 371–373.

[29] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Commun.*, vol. 35, no. 1–2, pp. 31–51, 2001.

[30] T. Niesler, "Language-dependent state clustering for multilingual acoustic modelling," *Speech Commun.*, vol. 49, no. 6, pp. 453–463, 2007.

[31] G. Heigold *et al.*, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8619–8623.

[32] N. T. Anh, H. T. T. Hang, and G. Chen, "One approach in the time domain in detecting copy-move of speech recordings with the similar magnitude," *Int. J. Eng. Appl. Sci.*, vol. 6, no. 4, 2019, doi: 10.31873/ijeas/6.4.2019.05.

[33] N. T. Anh, H. T. K. Dung, and N. T. P. Nhung, "Adaptive Cross-Correlation Compression Method in Lossless Audio Streaming Compression," *Int. J. Eng. Appl. Sci.*, vol. 6, no. 3, 2019, doi: 10.31873/ijeas/6.3.2019.30.

[34] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 7319–7323, doi: 10.1109/ICASSP.2013.6639084.

[35] S. Toshniwal *et al.*, "Multilingual speech recognition with a single end-to-end model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4904–4908.

[36] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of deep neural networks for multilingual speech recognition training and adaptation," in *Proc. of INTERSPEECH*, 2017, no. CONF.

[37] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based End-to-End Models for Small-Footprint Keyword Spotting," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, pp. 2037–2041, Mar. 2018, doi: 10.21437/Interspeech.2018-1777.

[38] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 2013, pp. 416–421, doi: 10.1109/ASRU.2013.6707766.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[40] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks," *Proc. 23rd Int. Conf. Mach. Learn.*, pp. 369–376, 2006, doi: 10.1145/1143844.1143891.

[41] Z. Zhang, D. Wang, and X. Zhang, "THCHS- 30 : A Free Chinese Speech Corpus," 2015.

[42] H.-T. Luong, H. Chi Minh City, and H.-Q. Vu, "A non-expert Kaldi recipe for Vietnamese Speech Recognition System," 2016.

[43] M. Sun *et al.*, "Max-Pooling Loss Training of Long Short-Term Memory Networks for Small-Footprint Keyword Spotting," *2016 IEEE Work. Spok. Lang. Technol. SLT 2016 - Proc.*, pp. 474–480, May 2017, doi: 10.1109/SLT.2016.7846306.

[44] B. S. Michiel Hermans, "Training and Analysing Deep Recurrent Neural Networks," vol. 1. 100AD.

[45] W. Song and J. Cai, "End-to-End Deep Neural Network for Automatic Speech Recognition," 2015.

[46] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017, doi: 10.1109/TNNLS.2016.2582924.

[47] Y. Wang and F. Tian, "Recurrent Residual Learning for Sequence Classification," 2016, pp. 938–943, doi: 10.18653/v1/d16-1093.

[48] Y. Zhuang, X. Chang, Y. Qian, and K. Yu, "Unrestricted vocabulary keyword spotting using LSTM-CTC," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016, vol. 08-12-Sept, pp. 938–942, doi: 10.21437/Interspeech.2016-753.

[49] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to Construct Deep Recurrent Neural Networks."

[50] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches."

Nguyen Tuan Anh, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, P.R. China 0084988086099.

Tran Thi Ngoc Linh, Faculty of Electronic Engineering, Thai nguyen University of Technology, Thai Nguyen, Vietnam, 0084945855155.

Dang Thi Hien, Faculty of Electronic Engineering, Thai nguyen University of Technology, Thai Nguyen, Vietnam, 0084983812903.