INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Improving wav2vec2-based Spoken Language Identification by Learning Phonological Features

Mostafa Shahin, Zheng Nan, Vidhyasaharan Sethu, Beena Ahmed

School of Electrical Engineering and Telecommunications, UNSW, Australia
{m.shahin, zheng.nan, v.sethu, beena.ahmed}@unsw.edu.au

Abstract

Spoken language identification (SLI) is a key component in speech-processing tools such as spoken language understanding. In code-switching conversational speech, speakers change languages for short durations posing an additional challenge to language identification techniques. In this work, we investigate the ability of a wav2vec2-based SLI method in identifying the spoken language of English/Mandarin code-switching child-directed conversational speech recorded via Zoom. The proposed system allows the pre-trained wav2vec2-based model to learn language-dependent phonological features by fine-tuning first on detecting manners and places of articulation, then on classifying between English and Mandarin speech segments. The proposed system was tested against parent-child Zoom recordings provided as a part of the MERLIon CCS challenge of language identification. The system achieved the best balanced accuracy of 81.3% and the second-lowest equal error rate of 10.6%.

Index Terms: language identification, speech attributes, wav2vec2, code-switching

1. Introduction

Spoken language identification (SLI), a core preprocessing phase of many multi-lingual speech systems (e.g., multi-lingual speech recognition, speech translation, speech transcription, etc.), identifies the language spoken in a given utterance [1, 2]. Current SLI techniques have explored the use of acoustic features such as linear predictive coding, filter banks, Mel frequency cepstral coefficients to identify the utterance language as they have been shown to present the dependency of phone units on languages [3, 4]. Prosodic and phonotactic features, which are not directly reflected by spectral features, have also been investigated as additional sources of knowledge since both levels of features provide complementary language cues [5, 6]. Having demonstrated effectiveness in speaker verification tasks, i-vector based approaches have also been successfully applied in SLI tasks [7, 8] and provided significant performance gain, with compact utterance-level i-vectors extracted to express distinctions among different languages [9]. To make the final language identification, these features have been fed to a number of different classifier types, with statistical models, particularly Gaussian mixture models and hidden Markov models, typically adopted [10–12].

All these approaches [3–12] are only effective for long utterances (typically no shorter than 3 sec). Their performance degrades drastically when utterance duration decreases [13]. However in real-world code-switching scenarios, utterances can be as short as 100 ms. Over the past decade, deep neural networks (DNNs) [14, 15], convolutional neural networks (CNNs) [16], recurrent neural networks (RNNs) [17, 18], and attention-based networks [19] have been successively used in SLI to better capture the temporal dependencies present in short utterances.

Given the lack of sufficient annotated data available to train SLI models, recent work has explored leveraging upon self-supervised speech representations learnt from large quantities of unannotated data via optimizing a contrastive predictive objective [20, 21], and then fine-tuning the model with limited annotated data for a downstream SLI task [22]. However, it has been shown that if the front-end of a model is first pre-trained with a relevant task (e.g., phoneme recognition), the bottleneck features extracted by this pre-trained front-end can be more effective for SLI than the raw acoustic features [13]. Since wav2vec2 is pre-trained to learn general purpose, mostly language-independent, speech representations, we envision that fine-tuning the pre-trained wav2vec2 first with a language-dependent task, can provide a better initialization point to fine-tune the ultimate SLI task.

In this paper, we thus investigate the effectiveness of pre-training the wav2vec2 framework with phonological features, specifically speech attributes. Speech attributes, such as the manners and places of articulations, provide a low-level description of sound production of the articulators involved and how these articulators move to produce a specific sound. Any spoken language and its phoneme can be characterized in terms of these attributes [23]. Here, we propose first fine-tuning the pre-trained wav2vec2 framework to learn language-dependent phonological features that include the manners and places of articulations, as well as features such as voiced. We use this approach to discriminate between English and Mandarin short utterances (1.1 sec on average) as a part of the MERLIon CCS challenge (language identification task).

The rest of the paper is organized as follows. Section 2 presents the proposed SLI system. Section 3 describes the experimental setup, while Section 4 provides the experimental results, followed by a conclusion in Section 5.

2. wav2vec2-based Spoken Language Identification

Fig. 1 details our proposed SLI system for short utterances. It consisted of two main blocks, a speech attribute detection network that feeds into a language identification network. The speech attribute detection network $\mathcal{M}_1$ used a pre-trained wav2vec2 architecture model followed by a linear layer in which the neurons represented N English speech attributes. The weights of the wav2vec2 part were initialized by a self-supervised pre-trained model $\mathcal{M}_0$, while the weights of the linear layer were randomly initialized. The input of the speech

10.21437/Interspeech.2023-2533

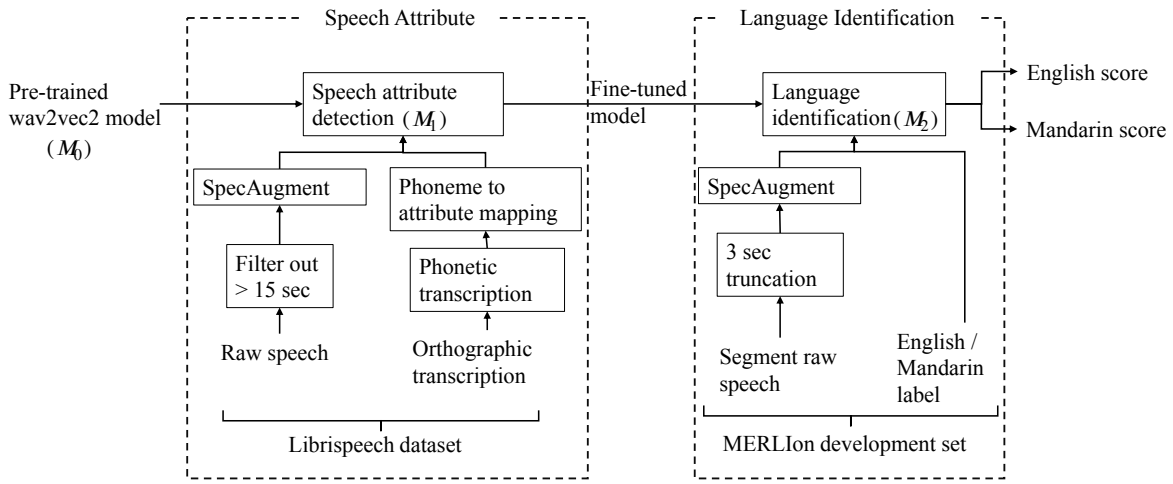

Figure 1: *Block diagram of proposed SLI system. The wav2vec2 pre-trained model is first fine-tuned using the Librispeech dataset to detect the existence or absence of a set of speech attributes. The resultant model is then fine-tuned to classify between English and Mandarin segments using the MERLIon challenge development set.*

attribute detection network $\mathcal{M}_1$ includes only raw Librispeech utterances of duration $<$ 15 sec, due to the limitations of the GPU memory size, and the target labels are N binary sequences corresponding to the N speech attributes for each utterance. The target sequences of each utterance were obtained by first converting the orthographic transcription into a phonetic transcription using a pronunciation dictionary [24] and then each phoneme was mapped to their corresponding speech attributes. A multi-label variant of the connectionist temporal classification (CTC) method was used as the loss function [25].

The language identification network $\mathcal{M}_2$ was obtained by replacing the last linear layer of $\mathcal{M}_1$ with a binary classification head that classifies each utterance as English or Mandarin. The wav2vec2 front-end of the network was first initialized using the parameters of the fine-tuned $\mathcal{M}_1$, before fed by raw speech segments of the MERLIon challenge development set with their corresponding language labels used as the target outputs. Because of the short-duration nature of the dataset, we truncated the training speech segments to 3 sec. The cross-entropy function was utilized to compute the loss of the language identification network. In both networks, the specaugment method [26] was used for data augmentation. Finally, the logit values of the binary classification layer were used as the utterance scores of English or Mandarin speech segments.

2.1. Speech Attribute Detection

Fig. 2 describes the speech attribute detection phase. As seen, the raw speech signal is first transformed into a sequence of speech representations using a pre-trained self-supervised speech representation model. Here we adopt the wav2vec2 framework [20]. The extracted speech representations are then fed to an output linear layer with dimension equal to the target number of classes.

Unlike phonetic representation where each acoustic unit is represented as a unique symbol or phoneme, phonological features are non-mutually exclusive therefore each phonological acoustic unit can have multiple phonological features. For instance, the phoneme */m/* is described as *nasal*, *voiced*, and *labial*. Therefore, the problem becomes a multi-label classification problem. Each speech utterance can thus be mapped to

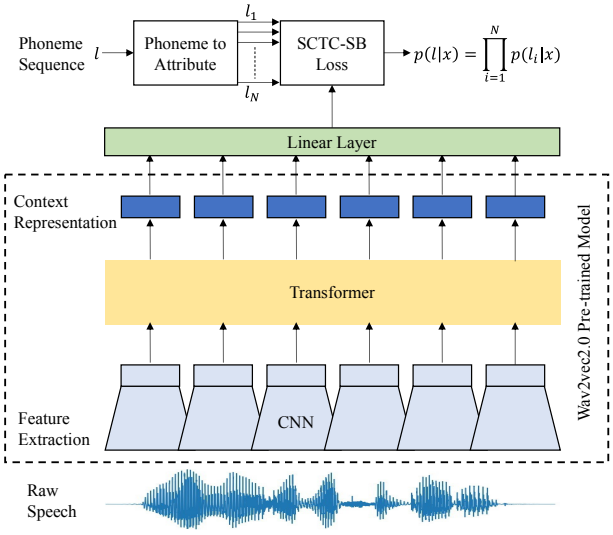

Figure 2: *Training procedure of the speech attribute detection model. A linear layer is added on top of a wav2vec2-based pre-trained model. For each speech utterance, the linear layer converts the corresponding phoneme sequence to N binary sequences of speech attributes. The CTC loss is then computed for each attribute sequence. Finally, the SCTC-SB loss is computed by multiplying all speech attributes' CTC losses.*

different attribute sequences.

As the standard CTC method only solves a single-label sequence-to-sequence problem [27], we used the separable connectionist temporal classification with shared blank (SCTC-SB) method, a multilabel variant of the CTC method we previously proposed in [25]. The SCTC-SB method works by computing the CTC loss over each labeling category separately and then adding them together (i.e., multiplying the conditional probabilities) to get the target loss and sharing the *blank* token among all categories. The SCTC-SB loss function was then estimated from the labeled data and the outputs of the linear layer. The gradients derived from the loss function were further backprop-

Table 1: *List of employed speech attributes.*

Manners	Places	Others
consonant, sonorant, fricative, nasal, stop, approximant, affricate, liquid, vowel, semivowel, continuant	alveolar, palatal, dental, glottal, labial, velar, mid, high, low, front, back, central, anterior, posterior, retroflex, bilabial, coronal, dorsal	long, short, monophthong, diphthong, round, voiced

Table 2: *The optimum training parameters of the system.*

Parameters	Speech Attribute	Language Identification
optimizer	AdamW	AdamW
loss	SCTC-SB	cross entropy
learning rate	1e-4	1e-4
weight decay	0.005	0.005
batch size	32	16
N# epochs	30	10
warmup steps	10% of total steps	10% of total steps

agated to fine-tune the wav2vec2 front-end.

We used 35 speech attributes representing the manners and places of articulation along with other phonological features as listed in Table 1. These attributes were selected so that each phoneme has a unique binary representation in terms of these 35 attributes. The 35 speech attributes were jointly learnt using the SCTC-SB criterion. A category for each attribute (att) was defined with items $C_i = \{+att, -att\}$. The category that represents the nasal attribute, for example, has possible outputs of +nasal,-nasal. Therefore, the number of network outputs is equal to 71, where 35 nodes represent the existence of each attribute ($+att$), 35 nodes represent the absence of each attribute ($-att$), and one node represents the blank output that is shared among all categories. The network was fine-tuned using 100 hours of the Librispeech dataset. The phonetic transcription of the dataset was extracted from the provided orthographic transcription using the CMU pronunciation dictionary [24]. Each phoneme was then mapped to $+att$, if the phoneme was characterized by the underlying attribute or $-att$ otherwise.

2.2. Language Identification

For the language identification phase, the wav2vec2 front-end of the language identification network $\mathcal{M}_2$ was initialized from $\mathcal{M}_1$, which has been fine-tuned for the speech attribute detection task as described in Section 2.1. Within $\mathcal{M}_2$, the last linear layer inherited from $\mathcal{M}_1$ was replaced by a randomly initialized classification head which consisted of a linear layer over the average pooled output of the contextual representation sequence. This linear layer has 2 dimensions, each of which outputs a score for a language (English/Mandarin). $\mathcal{M}_2$ was fine-tuned on a subset of the MERLIon CCS development set and tested on the rest of the development set. During the fine-tuning of $\mathcal{M}_2$, all speech utterances were truncated to 3 seconds. This forces the network to learn language discriminative features from short speech segments.

3. Experimental Setup

3.1. Datasets

Two datasets were employed in this work, namely, the Librispeech (LS) and MERLIon (MER) datasets. The LS dataset was used in the training and evaluation of the speech attribute detection model $\mathcal{M}_1$. The model was trained on 100 hours of the clean portion of the LS dataset as described in [28]. The development (LS_clean_dev) and testing (LS_clean_test) subsets of the clean portion were also utilized to monitor the training process and evaluate the model performance. The CMU dictionary was used to obtain the phonetic transcription of the LS orthographic transcription [29]. MER dataset was provided with the challenge and consisted of two subsets, the development (MER_dev) and evaluation (MER_eval) sets. Both MER_dev and MER_eval are composed of parent-child conversation recordings via Zoom. Each subset contains recordings from 56 different parent-child pairs of which $\sim$85% English and $\sim$15% Mandarin. The MER_dev set was associated with the language label of each segment while the MER_eval set contained only the speech segments and the performance is computed via CodaLab. Therefore, in this work, MER_dev set was further divided into training (MER_dev_train) and testing (MER_dev_test) subsets to train and compare the performance of different models. The MER_dev_train contained recordings from 46 parent-child pairs while the remaining 10 pairs formed the MER_dev_test set.

3.2. Training Procedure

The training of both speech attribute detection network $\mathcal{M}_1$ and the language identification network $\mathcal{M}_2$ was performed using the pytorch toolkit. Table 2 summarizes the training parameters used to achieve the best performance.

All speech data was resampled to 16 kHz sampling rate and the latent representations extracted from the feature extraction (i.e., CNN) layer of the wav2vec2 front-end were further augmented using SpecAugment [26]. In both tasks the feature extraction layer was frozen (i.e., not learnt) during the fine-tuning process.

3.3. Evaluation Metrics

For the speech attribute detection network, as the output is a binary sequence of $+att/-att$ symbols, the traditional error rate derived from the Levenshtein distance metric [30] was used. The Levenshtein distance metric works by measuring the difference between two sequences in terms of the number of insertion I, deletion D, and substitution S edits. Therefore, the attribute error rate (AER) is computed as:

$$\text{AER} = \frac{S + D + I}{N}. \tag{1}$$

For the language identification phase, two measures were considered, namely, the equal error rate (EER) and the balanced accuracy (BAC). The EER is defined as the point when both the false acceptance rate (FAR) and false rejection rate (FRR) are equal. In this challenge, FA occurs when a Mandarin segment is recognized as English and FR occurs when an English segment is recognized as Mandarin. The BAC is computed by first calculating the recall of the English and Mandarin classes separately and then averaging them.

3.4. Baseline Model

The provided baseline system is an end-to-end conformer model that is composed of four conformer encoder layers, followed by a statistics pooling layer and three linear layers with ReLU activation in the first two linear layers. Each self-attention encoder layer employs eight attention heads with input and output dimensions set at 512. Subsequently, the statistics pooling layer produces a 1024-dimensional output that is then projected onto the target languages using three linear layers of 1024, 512, and 2 nodes respectively. The model was trained on 200 hours of Aishell Mandarin data, 100 hours of Librispeech data, and 100 hours of National Speech Corpus data.

4. Experimental Results

Fig. 3 shows the performance of the speech attribute detection network $\mathcal{M}_1$ in terms of the AER of the 35 adopted attributes. The model was fine-tuned on the 100 hours LS_clean and tested against the LS_dev and LS_test sets. The figure shows that the accuracies of all speech attributes of both sets are above 99% (AER $< 1\%$).

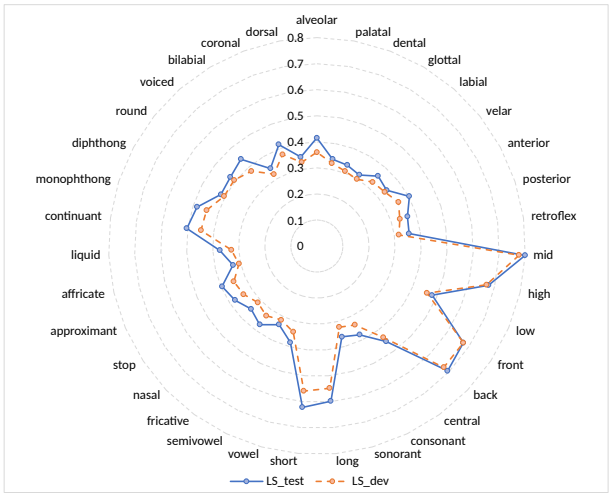

Figure 3: *The AER of the 35 speech attributes of the LS_dev and LS_test sets obtained when wav2vec2-large-robust pre-trained model was fine-tuned for speech attribute detection using SCTC-SB criteria.*

The performance of the proposed language identification model $\mathcal{M}_2$ and the challenge baseline model are summarized in Table 3. Our proposed model was trained by first fine-tuning the wav2vec2-large-robust model for the speech attribute task using the LS_clean dataset followed by fine-tuning for the language identification task using the MER_dev set. This system achieved the highest BAC of 81.3% and the second-lowest EER of 10.6% among all participants.

Table 3: *The performance of the language identification system.*

Model	**MER_dev_test**		**MER_eval**	
	EER	BAC	EER	BAC
Baseline	-	-	21.7%	50.9%
Proposed Model ($\mathcal{M}_2$)	10.65%	77.40%	10.6%	81.30%

To demonstrate the effectiveness of the speech attribute fine-tuning step on the language identification task, we conducted two experiments. First, we used the wav2vec2-large-robust pre-trained model and fine-tuned it directly to perform the language identification step by adding the classification head on top of the transformer output. Second, we fine-tuned the wav2vec2-large-robust model for a phoneme classification downstream task using LS_clean dataset, followed by fine-tuning for the language identification task. Table 4 compares the performance of the two models and the proposed speech attribute-based model in terms of their EER and BAC. The results show that fine-tuning of the wav2vec2 pre-trained model directly for the language identification task achieved EER and BAC of 11.3% and 76%, respectively. By fine-tuning the model first on phoneme recognition, the EER degraded slightly but the BAC improved from 76% to 77%. The lowest EER of 10.65% was obtained by fine-tuning the model first to detect the speech attributes followed by language classification.

Table 4: *Comparison of language identification performances of directly fine-tuned wav2vec2, wav2vec2 model that first fine-tuned on phoneme recognition task, and wav2vec2 model that first fine-tuned on attribute detection task.*

Model	**EER**	**BAC**
Wav2vec2	11.3%	76.0%
Wav2vec2 + Phoneme	11.4%	77.0%
Wav2vec2 + Attribute	**10.6%**	**77.3%**

In all previous models we froze the feature extraction part of the wav2vec2 model, i.e., the CNN feature encoder. Further training the feature encoder during the fine-tuning of the language identification task, slightly increased the EER from 10.6% to 10.8% while the BAC dropped from 77.3% to 76%.

5. Conclusion

This paper described the proposed SLI system submitted to the language identification task of the MERLIon CCS challenge. The system identified the language of short English and Mandarin utterances recorded from parent-child conversations via Zoom. Our proposed method was based on the wav2vec2-large-robust pre-trained model. The pre-trained model was first fine-tuned to recognize 35 speech attribute (phonological) features including manners and places of articulation. The speech attribute model was then fine-tuned on the language identification task to classify each utterance as English or Mandarin.

Experimental results showed that learning the language-dependent phonological features relatively improved the BAC and EER of the language identification task by $\sim$5% and $\sim$6%, respectively, compared to fine-tuning the pre-trained model directly for the language identification task. This is probably due to the phonological difference between the English and Mandarin languages. For instance, in English voiceness is a distinctive attribute that can distinguish one phoneme from another, such as /t/ and /d/, while it is not a distinctive attribute in Mandarin. Moreover, some phonological patterns such as consonant clusters, i.e. multiple consecutive consonants, are common in English while in Mandarin each consonant is followed by a vowel [31].

The proposed method achieved the highest BAC of 81.3% and the second lowest EER of 10.6% among all the participants of the language identification challenge.

6. References

[1] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[2] C.-H. Lee, "Principles of spoken language recognition," *Springer Handbook of Speech Processing*, pp. 785–796, 2008.

[3] J. Foil, "Language identification using noisy speech," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11. IEEE, 1986, pp. 861–864.

[4] M. Sarma and K. K. Sarma, "Long-term critical band energy-based feature set for dialect identification using a neuro-fuzzy approach," *IEEE Intelligent Systems*, vol. 33, no. 1, pp. 40–52, 2018.

[5] H. Li and B. Ma, "A phonotactic language model for spoken language identification," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 2005, pp. 515–522.

[6] R. Tong, B. Ma, D. Zhu, H. Li, and E. S. Chng, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[7] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[8] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011.

[9] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "i-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2013.

[10] M. A. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 397–406, 2012.

[11] K. Wong and M.-h. Siu, "Automatic language identification using discrete hidden markov model," in *Eighth International Conference on Spoken Language Processing*, 2004.

[12] M. A. Zissman, "Automatic language identification using gaussian mixture and hidden markov models," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1993, pp. 399–402.

[13] Z. Ma, H. Yu, W. Chen, and J. Guo, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE transactions on vehicular technology*, vol. 68, no. 1, pp. 121–128, 2018.

[14] I. Lopez-Moreno, J. Gonzalez-Dominguez, D. Martinez, O. Plchot, J. Gonzalez-Rodriguez, and P. J. Moreno, "On the use of deep feedforward neural networks for automatic language identification," *Computer Speech & Language*, vol. 40, pp. 46–59, 2016.

[15] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 5337–5341.

[16] A. Lozanodíez, R. Z. Candil, J. G. Domínguez, D. T. Toledano, and J. Gonzálezrodríguez, "An end-to-end approach to language identification in short utterances using convolutional neural networks," 2015.

[17] J. Gonzalez-Dominguez, I. Lopez-Moreno, and H. Sak, "Automatic language identification using long short-term memory recurrent neural networks," 2014.

[18] S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps, "Bidirectional modelling for short duration language identification," *Proc. Interspeech 2017*, pp. 2809–2813, 2017.

[19] W. Geng, W. Wang, Y. Zhao, X. Cai, and B. Xu, "End-to-end language identification using attention-based recurrent neural networks," *Interspeech 2016*, pp. 2944–2948, 2016.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[21] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *Advances in neural information processing systems*, vol. 32, 2019.

[22] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[23] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[24] C. M. University. (2014) Cmu pronouncing dictionary. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[25] M. Shahin, "Automatic screening of childhood speech sounds disorders and detection of associated pronunciation errors," Ph.D. dissertation, 2023.

[26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[29] R. Weide *et al.*, "The carnegie mellon pronouncing dictionary," *release 0.6, www. cs. cmu. edu*, 1998.

[30] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.

[31] F. Zhang and P. Yin, "A study of pronunciation problems of english learners in china," *Asian social science*, vol. 5, no. 6, pp. 141–146, 2009.