Interspeech 2022
18-22 September 2022, Incheon, Korea

PHO-LID: A Unified Model Incorporating Acoustic-Phonetic and Phonotactic Information for Language Identification

Hexin Liu[1], *Leibny Paola Garcia Perera*[2], *Andy W. H. Khong*[1], *Suzy J. Styles*[3], *Sanjeev Khudanpur*[2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]CLSP and HLT-COE, Johns Hopkins University, USA
[3]Psychology, School of Social Sciences, Nanyang Technological University, Singapore

HEXIN002@e.ntu.edu.sg, lgarci27@jhu.edu

Abstract

We propose a novel model to hierarchically incorporate phoneme and phonotactic information for language identification (LID) without requiring phoneme annotations for training. In this model, named PHO-LID, a self-supervised phoneme segmentation task and a LID task share a convolutional neural network (CNN) module, which encodes both *language* identity and sequential *phonemic information* in the input speech to generate an intermediate sequence of "phonotactic" embeddings. These embeddings are then fed into transformer encoder layers for utterance-level LID. We call this architecture CNN-Trans. We evaluate it on AP17-OLR data and the MLS14 set of NIST LRE 2017, and show that the PHO-LID model with multi-task optimization exhibits the highest LID performance among all models, achieving over 40% relative improvement in terms of average cost on AP17-OLR data compared to a CNN-Trans model optimized only for LID. The visualized confusion matrices imply that our proposed method achieves higher performance on languages of the same cluster in NIST LRE 2017 data than the CNN-Trans model. A comparison between predicted phoneme boundaries and corresponding audio spectrograms illustrates the leveraging of phoneme information for LID.

Index Terms: Language identification, acoustic phonetics, phonotactics, self-supervised learning, phoneme segmentation

1. Introduction

Spoken language identification (LID) refers to the process through which the language identity of a speech sample can automatically be determined [1]. Early studies in LID suggested that acoustic and phonotactic features are the most effective language cues [1, 2, 3]. In particular, acoustic features such as Mel-frequency cepstral coefficients (MFCCs) and filter bank energies are the most commonly used language cues in recent LID systems [4, 5, 6]. Existing acoustic LID models, in general, comprise a language encoder and a classifier. They can either be optimized separately such as the conventional i-vector and x-vector methods [5, 7], or integrated in an end-to-end neural network [8, 9]. In addition, the phoneme-aware acoustic LID method utilizes phonemic information to achieve high LID performance. Such acoustic-phonetic LID system is trained on acoustic features and incorporates phoneme information by jointly optimizing the LID task and a phoneme-related task such as automatic speech recognition and phoneme classification [9, 10, 11]. In existing phoneme-aware LID methods, the LID and phoneme-related tasks share the same language encoding module in which the language identities and phoneme information are learned. As a result, the phoneme-related branch requires phoneme annotations or transcriptions.

Phonotactics involves the permissible combinations of phonemes in languages. As opposed to an acoustic-phonetic unit that is shorter than a phoneme, a phonotactic unit can cover several phonemes. It is also useful to note that while phonemes can be shared across languages, the statistics of their sequential patterns differ from one language to another [1]. Conventional phonotactic LID methods comprise a phone recognition module followed by the language modeling for target languages [2, 12]. Typically, in the parallel phone recognition language modeling (PPRLM) LID model [2], a phone recognizer that can either be universal or language-specific is first trained. The language models then capture sequential patterns of the recognized phones before generating scores for the target languages.

Notwithstanding the above, a phonotactic LID system faces the same challenge as the phoneme-aware LID method—it requires phoneme annotations of speech samples during training. However, the process of annotation is usually time consuming, requires significant resources and domain expertise. In contrast, acoustic approaches require only digitized speech samples and their language labels; they have since gained popularity over the recent years.

Since acoustic-phonetic and phonotactic cues depict the language identities of a speech signal in different granularities, it is natural to consider both jointly to achieve high LID performance. The proposed phonetic and phonotactic LID (PHO-LID) method incorporates phonetic and phonotactic information hierarchically via a convolutional neural network-transformer encoder (CNN-Trans) without the use of phoneme annotations [13]. Inspired by the success of self-supervised phoneme segmentation in [14], the CNN module in our model is shared by the LID and self-supervised phoneme segmentation tasks, through which it learns the language and positional phoneme information during training. To perform phonotactic LID, our proposed method mimics the phone n-gram modeling in a statistical manner. The language and phoneme-aware output features of the CNN module are first aggregated by a statistics pooling layer for each short-duration (hundreds of milliseconds) segment of the input speech signal. The segment-level phonotactic embeddings are then generated before the utterance-level LID, in which a statistics pooling layer computes the utterance-level statistics.

Apart from the high LID performance resulting from the complementary characteristics of acoustic-phonetic and phonotactic features, our model possesses two desirable properties. Firstly, as opposed to existing phoneme or phonotactic-aware methods, our proposed approach does not require phoneme annotation or transcription due to the self-supervised training. Secondly, the proposed PHO-LID model combines different language cues in an end-to-end manner resulting in lower com-

10.21437/Interspeech.2022-354

plexity compared to the fusion of several subsystems with respect to different language cues [1].

2. Related work

In [14], the author proposed a self-supervised approach for phoneme segmentation. Given the representations of a raw speech signal, a CNN model is optimized to identify the spectral changes. This is achieved by minimizing a noise-contrastive estimation (NCE) loss [15, 16], in which the anchor, positive, and negative samples correspond to the current frame, its adjacent frames, and non-adjacent frames, respectively. Denoting $\mathrm{sim}(\cdot,\cdot)$ as cosine similarity between two vectors, the NCE loss for each frame $\mathbf{z}_i$ is computed via

$$\mathcal{L}(\mathbf{z}_i) = -\log \frac{e^{\mathrm{sim}\left(\mathbf{z}_i, \mathbf{z}_{i+1}\right)}}{\sum_{\mathbf{z}_j \in \left\{\mathbf{z}_{i+1}\right\} \cup D_M(\mathbf{z}_i)} e^{\mathrm{sim}\left(\mathbf{z}_j, \mathbf{z}_{j+1}\right)}}, \quad (1)$$

where $D_M(\mathbf{z}_i)$ is a set of M negative samples to $\mathbf{z}_i$. These negative samples are randomly selected from all non-adjacent frames. During inference, the similarities between every two adjacent frames are computed. The phoneme boundaries are then determined as where the similarity is lower than a predetermined threshold.

3. Methodology

3.1. PHO-LID model

As shown in Fig. 1, the proposed PHO-LID model incorporates both acoustic-phonetic and phonotactic information hierarchically and comprises two branches. To illustrate these two branches separately, we define $\mathbf{X} = (\mathbf{x}_t \in \mathbb{R}^{K\times F} | t = 1, ..., T)$ as input of the proposed model. Here, $\mathbf{X}$ comprises features extracted from the input speech signal partitioned in segments, and T is the number of segments. Each segment $\mathbf{x}_t$ includes K frames $[\mathbf{f}_{t,1}, ..., \mathbf{f}_{t,K}]^\intercal$, where $\mathbf{f}_{t,k}$ is an F-dimensional feature vector of the k-th frame in segment t, and the original speech signal consists of $T \times K$ frames. It is useful to note that, in general, the average duration of common phonemes varies between 50 ms to 200 ms. The number of frames in each segment is thus defined to allow that each phonotactic unit in the proposed model can cover minimum two phonemes. These partitioned features are fed into the CNN module, which is jointly optimized by the primary LID task and the self-supervised phoneme segmentation task over each segment $\mathbf{x}_t$.

3.2. Self-supervised phoneme segmentation

To extract phonotactic information, the model should assimilate phoneme information before capturing the sequential patterns across several frames. Similar to the method originally proposed in [14], we employ the self-supervised phoneme segmentation to supply the CNN module with phoneme information.

With reference to Fig. 1, the self-supervised phoneme segmentation branch consists of the shared CNN module and a linear layer. In [14], given an input raw audio, each output unit of their model corresponds to a 10 ms frame, which is of a higher granularity than common phonemes. The premise that adjacents frames belong to the same phoneme is therefore reasonable. As opposed to [14], the input features of our proposed model are of 25 ms receptive field with 20 ms step size similar to those in conventional LID methods. With a minimum phoneme length being approximately 50 ms, we adopt convolutional layers with a kernel size of 1 in the CNN module to achieve a shorter receptive field of the output unit than the minimum phoneme length. The output of the linear projection layer corresponding to the input segment-level features $\mathbf{x}_t$ is given by $\mathbf{Z} = (\mathbf{z}_i \in \mathbb{R}^G | i = 1, ..., K)$, where G is the output dimension of the phoneme segmentation branch. The utterance-level NCE loss is then computed via

$$L^{NCE} = \frac{1}{KT} \sum_{T} \sum_{\mathbf{z}_i \in \mathbf{x}_t}^{K} \mathcal{L}(\mathbf{z}_i), \quad (2)$$

where $\mathcal{L}(\mathbf{z}_i)$ denotes the frame-level NCE loss defined in (1).

As highlighted in Section 2, similarities between adjacent frames can reveal phoneme boundaries. Therefore, after optimization, positional phoneme information is intrinsically encoded in the phoneme segmentation branch. Since phonotactics characterizes the combinations of phonemes spanning a short proportion of speech, the learned phoneme information enables the proposed PHO-LID model to incorporate the phonotactic information in segment-level features.

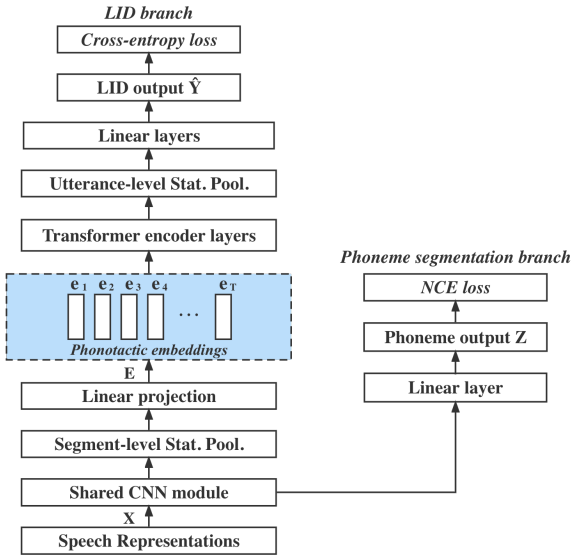

Figure 1: *The PHO-LID model.*

3.3. Supervised language identification

As the CNN module learns the phoneme and language information for each segment $\mathbf{x}_t$, the rest of the LID branch aims to perform phonotactic modeling. Features in each segment $\mathbf{x}_t$ are then aggregated as mean and standard deviation vectors, which are concatenated into a representation vector. These representations are linearly projected to segment-level phonotactic embeddings $\mathbf{E} = (\mathbf{e}_t \in \mathbb{R}^D | t = 1, ..., T)$, with each D-dimensional vector $\mathbf{e}_t$ corresponding to a segment $\mathbf{x}_t$. This statistical process serves as the phone n-gram modeling of the input speech. The transformer encoder layers subsequently capture the global dependencies of these phonotatic embeddings [13], and the output is aggregated into an utterance-level language embedding. Consequently, a score vector $\widehat{\mathbf{Y}}$ corresponding to C target languages is generated by the linear layers.

During training, we first pre-train the model on the self-supervised phoneme segmentation task for several epochs.Two strategies are employed and compared in the remaining updates. The first strategy solely updates the LID branch via a cross-

entropy loss, while the second strategy optimizes the model by a multi-task objective function. These strategies are, respectively, given by

$$L^{LID} = \text{CrossEntropy}\left(\widehat{\mathbf{Y}}, \mathbf{Y}\right), \quad (3)$$

$$L^{MUL} = \alpha L^{LID} + (1-\alpha) L^{NCE}, \quad (4)$$

where $\mathbf{Y}$ denotes the true language label and α is a parameter associated with multi-task learning. The model optimized by these two strategies are denoted as PHO-LID and PHO-LID-MUL, respectively. During inference, the input speech signal will be classified to the language of the highest score in $\widehat{\mathbf{Y}}$.

4. Data and model configuration

4.1. Datasets

We evaluated our approaches on the AP17-OLR and NIST LRE 2017 dataset. The AP17-OLR dataset contains training, development, and test sets [17]. The audios are of 16 kHz sampling rate and from ten oriental languages. We trained our systems on the training set without THCHS30 and evaluated them on the test set. The NIST LRE 2017 dataset, on the other hand, consists of the Fisher corpus [18], Switchboard corpus [19], a narrow-band telephony training set (TRN17) built from previous LRE data with over 2000 hours of audio data, a development set (DEV17), and an evaluation set (EVAL17) [20]. The DEV17 and EVAL17 comprise narrow-band MLS14 data and wide-band VAST data, while the MLS14 data consist of 3 s, 10 s, and 30 s duration and the VAST data comprise segments with speech duration ranging from 10 to 600 s. Fourteen languages exist in total within the data. We trained our systems on TRN17 and DEV17 and tested on the MLS14 data in EVAL17 to compare their performance on different duration levels.

4.2. Data preprocessing and feature extraction

The recordings in TRN17 of the MLS14 data in NIST LRE 2017 are first upsampled from 8 kHz to 16 kHz before an energy-based voice activity detection (VAD) is applied. The upsampled TRN17 recording are then partitioned into maximum 30 s segments before the feature extraction. The recordings in the AP17-OLR training set are partitioned into maximum 20 s segments without VAD before the feature extraction.

In this work, input features to the systems are wav2vec speech representations (W2V) extracted from the 16th context network block of the XLSR-53 cross-lingual wav2vec 2.0 model [21, 22]. This model has been pre-trained on fifty-three languages and 56,000 hours of speech data from Multilingual LibriSpeech, CommonVoice, and Babel [23, 24, 25]. The input wav2vec features $\mathbf{X}$ are of dimension $F = 1024$ and partitioned into segments, each of which comprises $K = 20$ frames before being fed into the model. Therefore, our phonotactic unit is approximately 400 ms long to cover minimum two phonemes. In addition, two baseline models in the experiments with respect to the NIST LRE 2017 data are trained on 80-dimensional bottleneck features (BN). The bottleneck features are extracted from an ASR model pre-trained on the Fisher and Swithchboard corpora using the Kaldi toolkit [26].

4.3. Model configuration

We provide two baseline x-vector and XSA-LID models for the experiments of NIST LRE 2017. With configurations following that of [7] and [27], respectively. The x-vector is trained by modifying the SRE16 recipe in the Kaldi toolkit with a back-end logistic regression classifier. The XSA-LID model is trained following the same strategy as the CNN-Trans model.

In our proposed model, the shared CNN module comprises three convolutional layers with kernel sizes $(1, 1, 1)$ and output dimensions $(512, 512, 512)$. The output dimension of the phoneme segmentation branch is given by $G = 64$. The outputs of the segment-level statistics pooling layer are projected to D-dimensional phonotactic embeddings with $D = 64$. The transformer encoder module consists of a layer norm followed by two transformer encoder layers, each of which has 8 heads, $d_{\text{model}} = 512$, and $d_{\text{ff}} = 2048$ [13]. The following linear layers comprise 512, 512, and C output nodes, respectively. The CNN-Trans model shares the same configuration as the PHO-LID model but excludes the phoneme segmentation branch.

The PHO-LID model is trained on the AP17-OLR data and NIST LRE 2017 data for 13 and 23 epochs, respectively. In the first 3 epochs, the CNN module is updated by the self-supervised phoneme segmentation task with a constant learning rate of 10^{-4}. In the remaining epochs, the model is updated with a learning rate that warms up from 0 to 1×10^{-4} in 3 epochs followed by the cosine annealing decay. The PHO-LID-MUL employ two α values 0.95 for AP17-OLR data and 0.97 for NIST LRE 2017 data. The Adam optimizer with batch size 128 is used. We evaluated our systems by employing accuracy (Acc.), equal error rate (EER) and the average cost (C_{avg}) [28]. The source code has been made available in GitHub. [1]

Table 1: *Evaluation on AP17-OLR by employing Accuracy (%), EER (%) and C_{avg}*

Method	Feat.	Acc.	EER	C_{avg}
i-vector [5, 17]	MFCCs	-	3.39	0.0352
PTN-LID [29, 17]	Fbanks	-	8.15	0.0689
CNN-Trans	W2V	96.74	1.10	0.0098
PHO-LID-3	W2V	97.57	0.80	0.0073
PHO-LID-10	W2V	97.67	0.82	0.0075
PHO-LID-MUL-3	W2V	**98.13**	**0.63**	**0.0058**
PHO-LID-MUL-10	W2V	97.50	0.86	0.0080

5. Experiment, results and analysis

5.1. Performance on AP17-OLR data

We present the evaluation results of the baseline and our proposed models on AP17-OLR data in Table 1. Results of the i-vector and phonetic temporal neural (PTN-LID) models are provided by the AP17-OLR evaluation plan [5, 29, 17]. The averaged results of the models across several trials are presented in Table 1. Here, the suffixes of PHO-LID and PHO-LID-MUL denote the number of negative samples M in (1). In Table 1, the PHO-LID-MUL-3 model achieves the highest LID performance compared to other methods and exhibits 40.82% relative improvement in terms of C_{avg} compared to the CNN-Trans model. This indicates that incorporating acoustic-phonetic and phonotactic information effectively enhances the LID performance. In addition, the PHO-LID-MUL-3 model achieves higher performance in terms of EER and accuracy compared to the PHO-LID-3. This suggests that the performance improvement is mainly attributed to the pre-training for self-supervised phoneme segmentation.

As M increases, the PHO-LID-MUL model suffers from significant performance degradation while model exhibits few variations of LID performance. Comparing the performance between the PHO-LID-MUL-3 and PHO-LID-3 models, these results imply that a small M with multi-task learning is preferred.

[1] https://github.com/Lhx94As/PHO-LID

Table 2: *Evaluation on the MLS14 set of NIST LRE 2017 by employing Accuracy (%), EER (%) and C_{avg} across three test speech durations 3 s, 10 s, and 30 s*

Method	Feat.	3s			10s			30s		
		Acc	**EER**	C_{avg}	**Acc**	**EER**	C_{avg}	**Acc**	**EER**	C_{avg}
x-vector [7]	BN	-	11.79	0.1159	-	7.81	0.0748	-	6.84	0.0654
XSA-LID [27]	BN	54.18	15.47	0.1685	74.90	7.39	0.0739	84.09	4.46	0.0406
	W2V	70.16	10.71	0.1004	84.32	4.89	0.0432	87.20	3.71	0.0329
CNN-Trans	W2V	**73.63**	**7.89**	0.0710	**86.36**	3.89	0.0346	89.93	2.91	0.0262
PHO-LID-3	W2V	72.61	8.22	0.0712	85.89	**3.84**	0.0344	**90.00**	**2.77**	**0.0242**
PHO-LID-MUL-3	W2V	72.52	8.10	**0.0696**	84.90	3.86	**0.0332**	89.81	2.87	0.0253

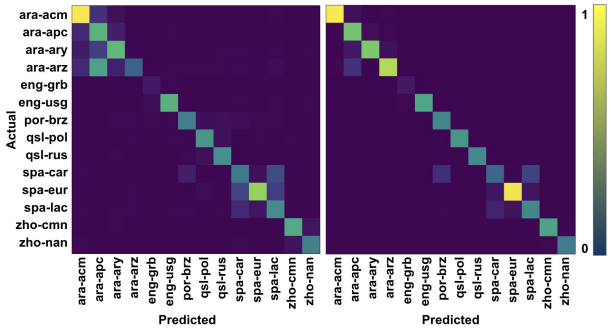

Figure 2: *Performance confusion matrices of the CNN-Trans model (left) and the PHO-LID-MUL-3 (right) on 30 s test speech in NIST LRE 2017 data.*

5.2. Performance on the MLS14 set of NIST LRE 2017

Table 2 presents the results on the MLS14 set of NIST LRE 2017 by the test speech durations. The proposed PHO-LID-MUL-3 model exhibits the highest overall performance compared to other systems in terms of C_{avg}, and both PHO-LID-3 and PHO-LID-MUL-3 models outperform the CNN-Trans model on overall performance in terms of C_{avg}. The above is consistent with the results shown in Table 1 and suggests the effectiveness of our proposed method. Moreover, compared to the CNN-Trans model, while the PHO-LID-MUL-3 achieves 1.97% relative improvement on 3 s test speech in terms of C_{avg}, it achieves 4.05% and 3.44% relative improvement on 10 s and 30 s speech, respectively. This is not surprising since phonotactic features are extracted from longer speech units as opposed to acoustic features.

We visualize the confusion matrices of the LID performance of the CNN-Trans and PHO-LID-MUL-3 models on 30 s test speech in Fig. 2. Compared to the CNN-Trans model, the proposed PHO-LID-MUL-3 model with phonemic and phonotactic information effectively reduces the confusion resulting from languages belonging to the same cluster—Arabic and Iberian clusters, and thus achieves higher performance.

In addition to the above, results presented in Tables 1 and 2 show that the W2V features introduce significant performance improvement compared to the use of conventional acoustic-phonetic features. This suggests that the W2V features are more discriminative and well-suited for the LID task. The comparison between these two tables also shows that the PHO-LID model trained on fewer data gains more improvement. This indicates that the phonotactic information, which intrinsically exists in W2V features, mitigates the lack of data.

6. Visulization and discussion

With reference to [14], phoneme boundaries occur where the similarity of two adjacent frames is lower than a predefined threshold. To demonstrate the existence of the positional phoneme information in our proposed model, we compare similarities between adjacent frames in a speech sample with the corresponding spectrogram. Since phoneme boundaries detection is not our target during inference, we annotate regions with significantly low similarities to indicate the predicted phoneme boundaries using dashed lines.

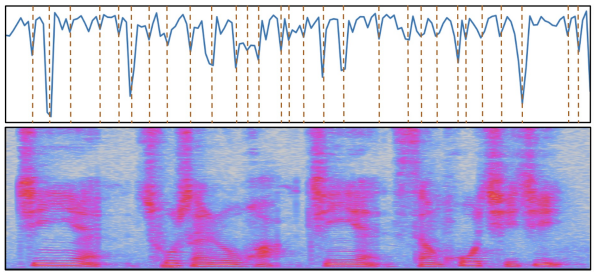

Figure 3: *Comparison between the phoneme boundaries predicted by our model (top) and the audio spectrogram (bottom).*

As shown in Fig. 3, the spectral changes can be approximately detected by our proposed model. This indicates that the model output contains positional phonetic information. Since our proposed phonotactic embedding is the concatenation of the mean and standard deviation vectors computed over frames in a segment, the above comparison highlights the feasibility of incorporating phonetic and phonotactic information.

7. Conclusion

We proposed the PHO-LID model that incorporates the phonetic and phonotactic information for LID without the need for phoneme annotations. Compared to the CNN-Trans model, our proposed model with multi-task optimization achieves higher performance on AP17-OLR and NIST LRE 2017 data, and the performance confusion matrices indicate that our proposed method can effectively distinguish languages of the same cluster in NIST LRE 2017. We visualize the predicted phoneme boundaries using the output units of the shared CNN module. The comparison between the predicted phoneme boundaries and the corresponding audio spectrogram shows the existence of phoneme information, which, in turn, highlights the feasibility of our proposed method.

8. Acknowledgements

This work was partially supported by the National Research Foundation, Singapore, under the Science of Learning programme (NRF2016-SOL002-011), the Centre for Research and Development in Learning (CRADLE) at Nanyang Technological University, and US the National Science Foundation via CCRI Award #2120435.

9. References

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

[2] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31–44, 1996.

[3] Y. Muthusamy, E. Barnard, and R. Cole, "Reviewing automatic language identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 33–41, 1994.

[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[5] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. Twelfth Annual Conf. Int. Speech Comm. Assoc.*, 2011.

[6] W. Cai, D. Cai, S. Huang, and M. Li, "Utterance-level end-to-end language identification using attention-based CNN-BLSTM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5991–5995.

[7] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Proc. Odyssey*, 2018, pp. 105–111.

[8] X. Miao, I. McLoughlin, and Y. Yan, "A new time-frequency attention mechanism for TDNN and CNN-LSTM-TDNN, with application to language identification," in *Proc. Interspeech*, 2019, pp. 4080–4084.

[9] S. Ling, J. Salazar, Y. Liu, and K. Kirchhoff, "Bertphone: Phonetically-aware encoder representations for utterance-level speaker and language recognition," in *Proc. Odyssey*, 2020, pp. 9–16.

[10] M. Zhao, R. Li, S. Yan, Z. Li, H. Lu, S. Xia, Q. Hong, and L. Li, "Phone-aware multi-task learning and length expanding for short-duration language recognition," in *Proc. APSIPA ASC*, 2019, pp. 433–437.

[11] Y. Liu, Z. Li, L. Li, and Q. Hong, "Phoneme-aware and channel-wise attentive learning for text dependent speaker verification," in *Proc. Interspeech*, 2021, pp. 101–105.

[12] R. Tong, B. Ma, H. Li, E. S. Chng, and K.-A. Lee, "Target-aware language models for spoken language recognition," in *Proc. Interspeech*, 2009, pp. 200–203.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[14] F. Kreuk, J. Keshet, and Y. Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," in *Proc. Interspeech*, 2020, pp. 3700–3704.

[15] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[16] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[17] Z. Tang, D. Wang, Y. Chen, and Q. Chen, "AP17-OLR challenge: Data, plan, and baseline," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf*, 2017, pp. 749–753.

[18] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *LREC*, vol. 4, 2004, pp. 69–71.

[19] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 1992, pp. 517–520.

[20] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, D. Reynolds, E. Singer, L. Mason, and J. Hernandez-Cordero, "The 2017 NIST language recognition evaluation," in *Proc. Odyssey*, 2018, pp. 82–89.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[22] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. Interspeech*, 2021, pp. 2426–2430.

[23] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. Interspeech*, 2020, pp. 2757–2761.

[24] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[25] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: BABEL project research at CUED," in *Proc. Int. Workshop Spoken Language Tech. for Under-resourced Languages*, 2014, pp. 16–23.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[27] H. Liu, L. P. G. Perera, A. W. H. Khong, J. Dauwels, S. J. Styles, and S. Khudanpur, "Enhance language identification using dual-mode model with knowledge distillation," *arXiv preprint arXiv:2203.03218*, 2022.

[28] A. F. Martin and A. N. Le, "NIST 2007 language recognition evaluation," in *Proc. Odyssey*, 2008, p. paper 16.

[29] Z. Tang, D. Wang, Y. Chen, L. Li, and A. Abel, "Phonetic temporal neural model for language identification," *IEEE/ACM Trans. Audio Speech Language Process.*, vol. 26, no. 1, pp. 134–144, 2017.