INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Language-universal phonetic encoder for low-resource speech recognition

Siyuan Feng, Ming Tu, Rui Xia, Chuanzeng Huang, Yuxuan Wang

Speech and Music Intelligence (SAMI), ByteDance

{fengsiyuan.ee,mingtu,rui.xia,huangchuanzeng,wangyuxuan.11}@bytedance.com

Abstract

Multilingual training is effective in improving low-resource ASR, which may partially be explained by phonetic representation sharing between languages. In end-to-end (E2E) ASR systems, graphemes are often used as basic modeling units, however graphemes may not be ideal for multilingual phonetic sharing. In this paper, we leverage International Phonetic Alphabet (IPA) based language-universal phonetic model to improve low-resource ASR performances, for the first time within the attention encoder-decoder architecture. We propose an adaptation method on the phonetic IPA model to further improve the proposed approach on extreme low-resource languages. Experiments carried out on the open-source MLS corpus and our internal databases show our approach outperforms baseline monolingual models and most state-of-the-art works. Our main approach and adaptation are effective on extremely low-resource languages, even within domain- and language-mismatched scenarios.

Index Terms: Language universal phonetic representation, multilingual training, low-resource

1. Introduction

Recently, end-to-end automatic speech recognition (ASR) systems have achieved remarkable performances in high-resource languages e.g. English [1, 2]. The availability of huge amounts of training data and the increase of computational capabilities are two essential driving forces. There are around 7000 languages in the world [3]. Most of the languages are considered under-resourced, i.e., having limited speech and/or text resources. This poses a major challenge in developing high-performance ASR systems for low-resource languages.

Low-resource ASR has been an active research topic recently [4, 5, 6, 7, 8, 9]. One mainstream research line is multilingual training [4, 5, 6], i.e. pooling together training speech and transcripts of multiple languages to train a single ASR model. A second research line is crosslingual transfer learning [7, 8], i.e., leveraging an ASR model trained by non-target, usually resource-rich language(s) as the seed model, retraining it with a target low-resource language's data. A third research line is self-supervised learning (SSL) [10, 1, 2, 11, 12], typically starting with unsupervised pretraining followed by supervised finetuning. SSL methods are effective especially when a large amount of untranscribed speech is available. In addition, data augmentation and self-training, which share the idea of increasing the supervised data amount, were studied for low-resource ASR [13, 14].

This paper follows the multilingual research line. The success of multilingual training for low-resource ASR suggests that the superiority of multilingual models over monolingual ones is presumably explained by (1) more training data; and (2) the ability of sharing linguistic (or more specifically, phonetic) knowledge across languages [15]. In this work, we are more interested in the second perspective. As different languages vary greatly in terms of fundamental units (phonemes), orthography, phonotactics etc, we argue that how to enable and facilitate phonetic representation sharing is not a trivial problem. In the era of hybrid deep neural network hidden Markov model (DNN-HMM) architecture, a predominant multilingual approach is sharing hidden layers between languages meanwhile keeping output layers language specific [4]. With this idea, phonetic information sharing is realized within hidden layers but not in output layers of the DNN model. In the E2E ASR paradigm, multilingual training is usually realized by directly pooling data of all the languages as the training data, and merging together graphemes of these languages as the model's vocabulary [6, 5, 16]. While the use of graphemes as basic modeling units brings simplicity and has been proven effective, it could suffer from the data sparsity issue especially in the multilingual setting [17]. Moreover, graphemes are not accurate descriptors of speech pronunciations in some languages (e.g. English), which might adversely affect multilingual phonetic sharing.

Past works explored basic modeling units other than graphemes for effective multilingual training in E2E ASR. In [17], the authors proposed to use bytes [18]. In [19], Huffman code was adopted. In [20, 21], subword units were adopted. The use of language-universal International Phonetic Alphabet (IPA) [22] symbols as basic modeling units was investigated in the E2E architectures including connectionist-temporal-classification (CTC) [23], E2E lattice-free maximum mutual information (LF-MMI) [24] and self-supervised pretraining [25]. On a different but relevant task of phone recognition, [15] analyzed the efficacy of IPA based basic units for multilingual phonetic sharing. They observed huge improvements of phone recognition from monolingual IPA models to the multilingual IPA model, especially on low-resource languages. Follow-up analyses [26, 27] compared model architectures between the listen-attend-spell (LAS) [28] and the DNN-HMM for IPA based phonetic sharing, and found that the LAS architecture is more effective.

From [15, 26, 27], it is clear that IPA is an effective means to facilitate multilingual phonetic representation sharing, particularly in the LAS model architecture. On the other hand, while there were studies on IPA based multilingual training in CTC and LF-MMI architectures [23, 24], the use of IPA in the LAS-based E2E models for ASR tasks has not been studied. Motivated by this, we attempt to leverage multilingual, hopefully language-universal phonetic representations learned by the LAS-based IPA model [15], to improve low-resource

10.21437/Interspeech.2023-621

ASR. Concisely, we propose to firstly train a multilingual IPA model with IPA transcribed speech, then finetune the IPA model with a target language's orthographically transcribed speech. An adaptation operation is optionally applied to the trained IPA model, in order to strengthen its phonetic representation learning of extremely low-resource languages, hence improve the ASR performance of these languages.

2. Proposed approach

The proposed approach consists of two main stages, which will be described in Sections 2.1 and 2.2 respectively. The optional adaptation method will be discussed in 2.1.1.

2.1. Language-universal IPA model

IPA is a standardized representation of speech sounds in written form [22]. It is independent of languages. This makes the IPA system intrinsically suitable for multilingual phonetic sharing. An IPA model can recognize a speech utterance into a sequence of speech sounds symbolized by the IPA system. To train a multilingual IPA model, N distinct languages' supervised training data (denoted as $\mathcal{F}$) is required. Grapheme-to-IPA conversion is applied to convert orthographic transcriptions of every language into phonetic IPA transcriptions. The vocabulary of the multilingual IPA model, i.e. the output layer of the model, consists of the union of IPA symbols present in the N languages. We treat modifier symbols, such as long vowels [ː] as separate basic units following [15]. For instance, the sound [aː] is recognized as two consecutive tokens [a], [ː]. This enables sharing between [a] and [aː]. Such sharing would not happen if [aː] were treated as a whole unit.

The multilingual IPA model learns a phonetic representation that is (quasi) language universal: fundamental speech sounds from different languages that share the same IPA symbol could be pronounced the same, or slightly differently, and they are mapped to the same output of the model. The fact that the language-universal IPA model captures a broad range of languages' phonetic information is preferred in subsequent ASR finetuning, especially for languages with limited data. Intuitively, increasing N and making languages in $\mathcal{F}$ more diverse help the IPA model approach language universality.

The LAS based E2E ASR architecture is adopted for developing the IPA model. The model consists of an encoder and a decoder. The language-universal phonetic representations are expected to be mainly modeled by the encoder of the IPA model.

2.1.1. Adaptation to an extremely low-resource language

In practice, it may well be that the language-universal IPA model's training data amounts of different languages are (highly) imbalanced. Phonetic representations of languages with relatively less data are at risk of being underfit by the IPA model, compared to phonetics of languages with more data. We hypothesize that this could result in sub-optimal ASR performance when finetuning the IPA model to an extremely low-resource language. To address this issue, we propose an adaptation operation on the IPA model. Given the unadapted, well-trained language-universal IPA model, targeting at a particular language, we utilize this language's IPA transcribed training data to retrain the model for a few epochs with a small learning rate. With such an operation, compared to the unadapted IPA model, the adapted IPA model is strengthened in capturing the phonetic representation of the target language, meanwhile without losing language universality.

2.2. Target-language ASR finetuning

A language-universal IPA model (either unadapted or adapted) is taken as the seed model. The IPA model's encoder is kept, while its decoder is replaced by a randomly initialized one of the same layer size and number. The seed model is finetuned with a target language's speech and orthographic transcripts. The vocabulary of the finetuned ASR model consists of the byte-pair encoding (BPE) tokens present in BPE-tokenized transcripts of the target language.

In principle, ASR finetuning could also be done on the whole language-universal IPA model, instead of only on the encoder part. However, the decoder of the LAS-based IPA model heavily captures phonotactics information of the training languages, and according to [26], phonotactics of non-target languages would hurt the recognition performance on a target language. Therefore, in this work we focus on finetuning only the encoder of the language-universal IPA model.

3. Experimental setup

3.1. Databases and evaluation metric

Speech corpora used for training the language-universal IPA models include the open-source Multilingual Speech (MLS) corpus [29] and our internal data. The MLS corpus covers Polish (PL), Portuguese (PT), Italian (IT), Spanish (SP), French (FR), Dutch (DU), German (GE) and English (EN), all derived from read audiobooks. The amounts of training, development and test data per language are listed in Table 1. Our internal data consists of EN and Japanese (JP), both derived from the video domain. The total hours of EN and JP internal data are 12.3k and 8.8k respectively. We consider the following multilingual training sets for training language-universal IPA models: **MLS-7**: we merge training sets of PL, PT, IT, SP, FR, DU and GE, summing up to **6.0k** hours; **MLS-8**: we merge **MLS-7** with the EN training set, summing up to **50k** hours. **NT-2**: we merge our EN and JP internal datasets, summing up to **21.1k** hours.

Data used for target-language ASR finetuning is taken from MLS only. As this paper focuses primarily on low-resource ASR, and the EN training set size in the MLS corpus is an order of magnitude larger than the other languages' training sets, we do not carry out English ASR finetuning. For each of the 7 MLS languages, the full training set and a training subset of 100 hours are prepared for finetuning respectively. The 100-hour training subset is randomly chosen from the full training set of a language.

The finetuned ASR model for a target language is evaluated on the language's test set in the MLS corpus, using word error rate (WER) as the metric. Throughout our experiments, the development data partition is used exclusively for monitoring the training process.

The relations of the multilingual training sets and MLS test sets are summarized in Table 2. The column 'Domain matched' denotes if the training set and the test sets are from the same domain, and the column 'Languages' denotes the existence of training languages covered (target, T) and not covered (non-target, N) by the test sets.

3.2. Language-universal IPA model

An open-source LanguageNet software [30] is utilized to convert orthographic transcripts of all the languages used in the experiments into phonetic IPA transcripts. The number of IPA symbols covered by every language in the MLS corpus is listed

Table 1: *The sizes of the MLS training, development and test sets in hours for every language, and the number of IPA symbols covered by every language.*

Language	PL	PT	IT	SP	FR	DU	GE	EN
Train	104	161	247	918	1.1k	1.6k	2.0k	44.7k
Dev	2.1	3.6	5.2	10.0	10.1	12.8	14.3	15.8
Test	2.1	3.7	5.3	10.0	10.1	12.8	14.3	15.6
# IPA symbols	35	37	29	31	45	40	47	48

Table 2: *Summary of domain match/mismatch between training and test data, and existence of target/non-target languages in the training data.*

Training set	Domain matched	Languages
MLS-7	Y	T
MLS-8	Y	T&N
NT-2	N	N
MLS-7&NT-2	Y&N	T&N

in Table 1. The numbers of IPA symbols covered by the *MLS-7* and *MLS-8* sets, i.e. union sizes of IPA symbols covered by the multiple languages, are 87 and 95 respectively. For the EN and JP internal data, their numbers of IPA symbols are 48 and 22. The numbers of IPA symbols covered by *NT-2* and *MLS-7&NT-2* are 55 and 96.

The IPA model consists of a Conformer [31] encoder and a Transformer [32] decoder. It is implemented using ESPnet [33]. The encoder and decoder have 18 and 2 layers respectively, with 4 attention heads, 768 attention dimensions and 2048 position-wise feed forward (PFF) dimensions. The encoder contains a 2-layer CNN with a kernel size of 31. Multiple IPA models are trained, each using one of the multilingual training sets in $\{$MLS-7, MLS-8, NT-2, MLS-7&NT-2$\}$. In every training experiment, the model is trained for 60 epochs with joint CTC and attention objectives and the CTC weight is 0.1, using the Adam optimizer [34], a peak learning rate of 0.001 and a warm-up step size of 2000. The final model is obtained by averaging over models of epochs 51 to 60.

3.3. IPA model adaptation

Adaptation is performed on the language-universal IPA model trained using the MLS-8 training set (denoted as the *unadapted IPA model*). Adaptation is carried out once for a target language in the MLS corpus (excluding EN). The unadapted IPA model is retrained with the learning rate of 5×10^{-5} for a fixed 2 epochs. The the learning rate and epochs are determined based on our pilot experiments.

3.4. Target-language ASR finetuning

ASR finetuning is performed using one language's training set in MLS (excluding EN). ASR finetuning used BPE-tokenized orthographic transcripts. The BPE tokenization was implemented by SentencePiece[1]. The number of BPE tokens for each language is 5000. Prior to finetuning, the IPA model's decoder is randomly initialized, and its output layer is replaced with a new one, which corresponds to the BPE tokens present in the target language. The number of training epochs, learning rate, warm-up size, objective function and model averaging follow the settings in IPA model training (see Section 3.2). Throughout this work, a language model is not used.

Table 3: *WER% of the proposed approach, baseline monolingual models and state of the arts on MLS test sets. The number enclosed in brackets denotes the number of parameters in a model.*

	PL	PT	IT	SP	FR	DU	GE	Avg.
MLS monolingual [29]	21.66	20.52	11.78	6.68	6.58	13.09	7.10	12.49
+ 5-gram LM [29]	20.39	19.49	10.54	6.07	5.58	12.02	6.49	11.51
XLSR-53 (300M) + 4-gram LM [9]	17.2	14.7	10.4	6.3	7.6	10.8	7.0	10.6
B0 (15 lang. init.; 370M) [6]	10.9	15.5	10.1	4.7	6.1	11.1	5.0	9.1
E3 (15 lang. init.; 1B) [6]	10.4	15.2	8.8	4.2	4.9	9.9	4.3	8.2
JUST (600M) [35]	6.6	8.0	8.2	3.7	5.2	9.5	4.1	6.5
Proposed approach and monolingual baseline finetuned with full training data								
Baseline monolingual (216M)	15.93	24.85	14.00	6.25	6.02	12.86	7.25	12.45
MLS-8 (216M)	6.98	12.71	10.43	4.84	4.51	10.80	6.34	8.09
+ Adaptation	6.84	12.48	10.39	5.04	4.69	10.82	6.34	8.09
MLS-7 (216M)	14.14	14.84	10.89	5.36	4.97	11.37	6.90	9.78
NT-2 (216M)	9.56	16.80	11.77	6.74	5.12	12.49	7.00	9.93
MLS-7&NT-2 (216M)	7.38	13.64	10.53	5.00	4.91	11.44	6.60	8.50
Proposed approach and XLSR-53 finetuned with 100-hour training data								
XLSR-53 (300M) + 4-gram LM [9]	18.9	15.7	12.0	7.9	9.8	10.9	7.4	11.8
MLS-8 (216M)	7.00	13.17	11.93	7.96	7.58	14.38	9.49	10.22
MLS-7 (216M)	16.41	15.54	12.00	7.97	8.62	15.62	10.45	12.37

3.5. Baseline monolingual ASR

Monolingual ASR models are trained and serve as the baseline. All the monolingual models take the same architecture as the ASR model described in Section 3.4. For every target language in the MLS corpus (excluding EN), speech and BPE-tokenized transcripts of the training set are used to train a monolingual model from scratch. The implementation of BPE tokenization keeps the same as that in Section 3.4. The number of training epochs, learning rate, warm-up size, objective function and model averaging follow the settings in Sections 3.2 and 3.4. In other words, a baseline monolingual ASR model for a target language and the proposed model finetuned from the language-universal IPA model to the target language differ only in the initialization of the model's encoder.

4. Results and discussion

4.1. Main results

Word error rate (WER) results of the proposed approach, the baseline monolingual models and state-of-the-art works on the MLS test sets are listed in Table 3. The proposed approach, irrespective of using any of the *MLS-7*, *MLS-8* and *NT-2* multilingual training sets for training the IPA model, performs significantly better than the baseline monolingual models and the MLS official monolingual models [29] in terms of the averaged WER. Notably, while the monolingual models by [29] and by ours are implemented differently (wav2letter++ [36] v.s. ESPnet), with different training objectives (CTC v.s. hybrid CTC/attention), the averaged WER numbers are close (12.49% v.s. 12.45%). Our best system, i.e. finetuning from the IPA model trained by the MLS-8 set, achieves 4.4% absolute WER reduction compared to the baseline monolingual system. Looking at WER break-down to each target language, a positive correlation is found between the absolute WER reduction (from monolingual to multilingual) of a language and the amount of training data of that language: Polish and Portuguese benefit the most and German has the smallest WER improvement. The results demonstrate the effectiveness of language-universal phonetic representations learned by the IPA model encoder for improving ASR performances, particularly for languages with very limited training data.

Table 3 shows that, the best result achieved by the proposed approach performs better than XLSR-53 [9], B0 and E3 [6]. Our approach does not perform better than the JUST [35] ex-

[1] https://github.com/google/sentencepiece

cept on the French set. The XLSR-53 utilizes 56k hours of unsupervised data (including but not limted to MLS) for pretraining and a 5-gram LM during decoding. The B0 and E3 models use 359k hours of training data covering 15 languages to train a seed model, followed by using the MLS training data to finetune the ASR model. E3 has a larger model size than all the other models in Table 3. The JUST model uses the MLS full training sets for joint unsupervised and supervised training. The superior results of JUST seem to suggest that the integration of self-supervised and supervised training strategies is a promising research direction, which we leave for future study.

Table 3 lists WER results of the proposed approach and XLSR-53 by using only 100 hours of data per language in ASR finetuning. It should be noted that the 100-hour finetuning sets used for XLSR-53 are not identical to the sets used in our experiments, since the 100-hour subset information could not be found in [9]. In the 100-hour finetuning scenario, our approach taking the IPA model trained with the MLS-8 set performs better than XLSR-53.

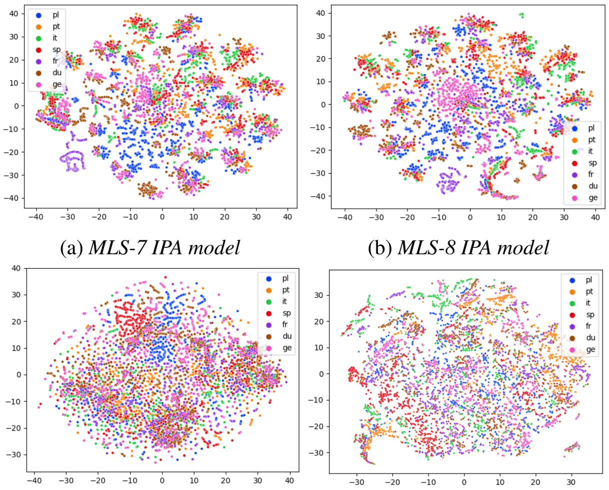
4.2. Impact of adaptation

WER results of the proposed approach by applying adaptation are listed in Table 3, in the row "+ Adaptation". The adaptation operation contributes to WER improvements on Polish, Portuguese and Italian, the three languages containing the smallest amounts of training data (100 $\sim$ 250 hours). For German and Dutch, the two languages having the largest amounts of training data, adaptation results in little to no difference. For Spanish and French, adaptation leads to WER degradation. Based on the WER results, we conclude that the proposed adaptation operation has a positive impact on target languages with extremely low-resource languages e.g. less than 250 hours. Suggested by the results, for languages with more training data, the proposed adaptation is expected to be less or not necessary.

4.3. Discussion about non-target languages and domain mismatch

In the proposed approach, the comparison between adopting MLS-8 and MLS-7 as the IPA model training data in Table 3 shows that including the non-target language (English) in training a language-universal IPA model is helpful to target languages' ASR tasks. Moreover, this benefit is seen on all the seven target languages. Thus a non-target language's phonetic knowledge information is helpful in the proposed approach.

The system by adopting the NT-2 set for IPA model training achieves an averaged WER better than the baseline monolingual models (see Table 3). As the NT-2 training set is from a different domain than that of the MLS test sets, and is composed of non-target languages only (English and Japanese), the results indicate the proposed approach is effective even on the domain- and language-mismatched scenario. On the other hand, the system with the NT-2 training set performs worse than the systems with MLS-7, MLS-8 and MLS-7&NT-2. This is in line with expectation: MLS-7, MLS-8 and MLS-7&NT-2 all contain domain-matched data of target languages. We observe an improvement by using the MLS-7&NT-2 training set in IPA model training, comparing to the systems with MLS-7 and with NT-2. This can be explained by a more language-universal phonetic representation learned in the system MLS-7&NT-2. Lastly, the system MLS-7&NT-2 underperforms the system MLS-8, which is probably due to the domain mismatch issue.



(a) *MLS-7 IPA model*

(b) *MLS-8 IPA model*

(c) *NT-2 IPA model*

(d) *XLSR-53*

Figure 1: *T-SNE visualizations on encoder representations of IPA models and the XLSR-53 [9] Transformer representation.*

4.4. Visualization of IPA model encoder representations

To gain a deeper understanding on the phonetic representations learned by the IPA model, we apply t-SNE [37] on the encoder output of several IPA models, and illustrate visualization results in Figures 1a, 1b and 1c. An illustration result on the open-source XLSR-53's Transformer output representation is shown in Figure 1d. In one figure, every sample point stands for a speech frame, and the color denotes the language. These speech frames are randomly chosen from the MLS test sets, 1000 frames per language. The selected frames are fixed for all the IPA models and the XLSR-53. From Figures 1a and 1b we observe distinct phone-like patterns consisting of samples in different colors. This indicates the two language-universal IPA models, MLS-7 and MLS-8, are able to group together speech sounds of different languages that share the same or similar pronunciations. In comparison, from Figure 1d, the representation learned by XLSR-53 is much less evident on phone-like patterns. From Figure 1c, we observe that for the IPA model trained with the domain- and language-mismatched NT-2 set, clusters of samples in different colors exist to a mild extent.

5. Conclusions

This paper presented an approach to improving low-resource ASR, which leverages language-universal phonetic representations learned by an IPA model. An optional adaptation operation was proposed on the IPA model, to strengthen the IPA model's ability of capturing extremely low-resource languages' phonetic representations. Experiments carried out on the MLS corpus and our internal databases showed the proposed approach outperforms baseline monolingual models and most of state-of-the-art works. Our main approach and adaptation are effective particularly on extremely low-resource languages. The approach brings improvements to the monolingual baseline even with domain- and language-mismatched training data. Visualizations of IPA models' learned representations further confirmed the IPA model is capable of capturing language-universal phonetic representations.

6. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 3451–3460, 2021.

[3] O. Scharenborg, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang *et al.*, "Speech technology for unwritten languages," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 964–975, 2020.

[4] 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 *Proc. ICASSP*, 2013, pp. 7304–7308.

[5] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP*, 2018, pp. 4904–4908.

[6] B. Li, R. Pang, T. N. Sainath, A. Gulati, Y. Zhang, J. Qin, P. Haghani, W. R. Huang, M. Ma, and J. Bai, "Scaling end-to-end models for large-scale multilingual asr," in *Proc. ASRU*, 2021, pp. 1011–1018.

[7] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. SLT*, 2012, pp. 246–251.

[8] W. Hou, H. Zhu, Y. Wang, J. Wang, T. Qin, R. Xu, and T. Shinozaki, "Exploiting adapters for cross-lingual low-resource speech recognition," *IEEE/ACM Trans. ASLP*, vol. 30, pp. 317–329, 2021.

[9] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint*, 2020.

[10] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[11] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. ASRU*, 2021, pp. 244–250.

[12] A. Baevski, W. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint*, 2022.

[13] L. Meng, J. Xu, X. Tan, J. Wang, T. Qin, and B. Xu, "Mixspeech: Data augmentation for low-resource automatic speech recognition," in *Proc. ICASSP*, 2021, pp. 7008–7012.

[14] Z.-Q. Zhang, Y. Song, M.-H. Wu, X. Fang, and L.-R. Dai, "Xlst: Cross-lingual self-training to learn multilingual representation for low resource speech recognition," *arXiv preprint arXiv:2103.08207*, 2021.

[15] P. Żelasko, L. Moro-Velázquez, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "That Sounds Familiar: An Analysis of Phonetic Representations Transfer Across Languages," in *Proc. INTERSPEECH*, 2020, pp. 3705–3709.

[16] S. Kim and M. L. Seltzer, "Towards language-universal end-to-end speech recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 4914–4918.

[17] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *Proc. ICASSP*, 2019, pp. 5621–5625.

[18] D. Gillick, C. Brunk, O. Vinyals, and A. Subramanya, "Multilingual language processing from bytes," *arXiv preprint arXiv:1512.00103*, 2015.

[19] Q. Liu, Y. Yang, Z. Gong, S. Li, C. Ding, N. Minematsu, H. Huang, F. Cheng, and S. Kurohashi, "Hierarchical softmax for end-to-end low-resource multilingual speech recognition," *arXiv preprint arXiv:2204.03855*, 2022.

[20] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *arXiv preprint arXiv:2007.03001*, 2020.

[21] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," *arXiv preprint arXiv:1806.05059*, 2018.

[22] I. P. Association, I. P. A. Staff *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[23] S. Tong, P. N. Garner, and H. Bourlard, "Multilingual training and cross-lingual adaptation on ctc-based acoustic model," *arXiv preprint arXiv:1711.10025*, 2017.

[24] ——, "An investigation of multilingual asr using end-to-end lf-mmi," in *Proc. ICASSP*. IEEE, 2019, pp. 6061–6065.

[25] S. Feng, M. Tu, R. Xia, C. Huang, and Y. Wang, "Language-universal phonetic representation in multilingual speech pretraining for low-resource speech recognition," in *Proc. INTERSPEECH (to appear)*, 2023.

[26] S. Feng, P. Żelasko, L. Moro-Velázquez, A. Abavisani, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "How phonotactics affect multilingual and zero-shot asr performance," in *Proc. ICASSP*, 2021, pp. 7238–7242.

[27] P. Żelasko, S. Feng, L. M. Velázquez, A. Abavisani, S. Bhati, O. Scharenborg, M. Hasegawa-Johnson, and N. Dehak, "Discovering phonetic inventories with crosslingual automatic speech recognition," *Computer Speech & Language*, vol. 74, p. 101358, 2022.

[28] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016, pp. 4960–4964.

[29] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[30] M. Hasegawa-Johnson, L. Rolston, C. Goudeseune, G.-A. Levow, and K. Kirchhoff, "Grapheme-to-phoneme transduction for cross-language ASR," in *Proc. ICSLSP*. Springer, 2020, pp. 3–19.

[31] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *Proc. ICASSP*, 2021, pp. 5874–5878.

[32] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs rnn in speech applications," in *Proc. ASRU*, 2019, pp. 449–456.

[33] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.-E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *Proc. INTERSPEECH*, pp. 2207–2211, 2018.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, vol. abs/1412.6980, 2014.

[35] J. Bai, B. Li, Y. Zhang, A. Bapna, N. Siddhartha, K. C. Sim, and T. N. Sainath, "Joint unsupervised and supervised training for multilingual asr," in *Proc. ICASSP*, 2022, pp. 6402–6406.

[36] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "Wav2letter++: A fast open-source speech recognition system," in *Proc. ICASSP*, 2019, pp. 6460–6464.

[37] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.