Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Switch Conformer with Universal Phonetic Experts for Multilingual ASR

Masato Mimura[1], *Jaeyoung Lee*[2], *Tatsuya Kawahara*[2]

[1]NTT, Inc., Japan
[2]School of Informatics, Kyoto University, Japan

email@address

Abstract

Multilingual end-to-end ASR presents significant challenges due to the need to accommodate diverse writing systems, lexicons, and grammatical structures. Existing methods often rely on large models with high computational costs for adequate cross-language performance. To address this, we propose the switch Conformer, which enhances model capacity while maintaining nearly the same inference cost as a standard Conformer. Our approach replaces the FFN module in each Conformer block with a sparse mixture of independent experts, activating only one expert per input to enable efficient language-specific feature learning. In addition, a shared expert trained with phonetic supervision captures language-universal speech characteristics. Experiments on streaming ASR using the CommonVoice dataset demonstrate that these experts work synergistically to achieve better performance than the baseline Conformer, with minimal additional active parameters.

Index Terms: multilingual speech recognition, Conformer, mixture of experts, IPA, articulatory features

1. Introduction

The performance of end-to-end (e2e) automatic speech recognition (ASR) [1, 2, 3, 4] has significantly improved primarily due to advancements in encoder design [5, 6] and data augmentation techniques [7, 8]. Furthermore, practically important but data-scarce tasks, such as multilingual or noise-robust ASR, have greatly benefited from recent developments in self-supervised or weakly-supervised training strategies [9, 10, 11, 12, 13, 14].

Multilingual e2e ASR enables a single model to transcribe speech across multiple languages, eliminating the need for separate models per language and improving performance in low-resource languages by leveraging data from high-resource ones [15, 16]. However, as shown in models like Whisper [14] and XLSR [13], it often requires large architectures with high computational costs, limiting their feasibility for on-device or streaming ASR. Furthermore, joint training across languages without considering linguistic similarity can lead to negative knowledge transfer, degrading performance [16, 17].

To address these issues, we propose the *switch Conformer* architecture, augmented with shared *universal phonetic experts*. We first modifies the original Conformer model [18] by replacing its feed-forward network (FFN) module with a sparse mixture of independent experts (SMoE). This largely increases the capacity and expressiveness of the model, enabling it to better accommodate various languages. Importantly, this modification maintains almost the same inference cost as the standard Conformer per frame, as each input is selectively routed to a single expert with the same size as the Conformer FFN.

Then, to facilitate positive knowledge transfer across languages, we also introduce a small expert to capture language-independent speech characteristics. This expert is shared among all inputs, unlike those in the SMoE. Here, we specifically focus on universal phonetic information, which can be systematically represented using symbols from international phonetic alphabet (IPA). By integrating this phonetic expert with the SMoE, our framework aims to balance model capacity and computational efficiency, and provide an effective inductive bias for multilingual training on linguistically diverse languages.

2. Preliminaries

2.1. Multilingual ASR

Multilingual e2e ASR enables speech recognition across multiple languages using a single deep sequence model. It is implemented through various e2e frameworks [2, 3, 4, 1]. The output layer's vocabulary can be either a union of all grapheme units across involved languages [15] or more universal byte-level tokens [14]. This approach eliminates the need for separate models per language and, more importantly, significantly improves recognition performance for low-resource languages due to data augmentation via high-resource languages.

Recent advancements in self-supervised and weakly supervised learning have enabled the use of large-scale real-world data, fundamentally improving multilingual ASR performance [13, 14]. However, these methods often require large models, which leads to high computational costs during inference. This is likely due to the need to accommodate variations in writing scripts, lexicons, and grammatical structures.

Another issue in the current framework of multilingual ASR is that simply increasing the number of languages does not necessarily improve performance for each one [16, 17]. In particular, within the capacity constraints of typical ASR models, training linguistically unrelated or geographically distant languages together can lead to degraded performance, such as Cyrillic-script and Quechuan languages [16].

2.2. IPA

E2e ASR generally uses grapheme-based tokens such as subwords and characters, as output units. However, these units are inherently tied to a language's writing system and cannot be easily shared across languages. As a result, models trained solely on grapheme targets struggle to transfer knowledge from one language to another.

To address this, incorporating lower-level, language-independent phonetic units as auxiliary targets can improve joint training efficiency. For example, Adams et al. [16] showed that using an auxiliary phoneme prediction task improves over-

10.21437/Interspeech.2025-1359

all performance. The international phonetic alphabet (IPA) offers a standardized system for transcribing speech sounds across languages, making it a useful representation of language-universal phonemes. By explicitly encoding distinct phonetic features, IPA-based targets in multilingual training can better capture cross-linguistic characteristics in speech.

Several grapheme-to-phoneme (g2p) tools convert grapheme transcriptions to IPA sequences [19, 20, 21]. However, most are dictionary-based and do not account for context-dependent phonological variations, leading to inevitable errors. For example, the ByT5-based multilingual g2p tool [19] achieves a phone error rate below 5% for half of its supported languages but can reach up to 30% for low-resource languages. To mitigate g2p errors, IPA prediction should be used as an auxiliary task with a lower importance weight [16, 17] or supplemented with more universal articulatory features [22, 23, 17].

2.3. Sparse mixture of experts

The mixture-of-experts (MoE) framework was originally proposed to enhance model expressiveness [24]. It consists of multiple subnetworks, each specialized in handling a subset of the complete set of training examples.

A key variant, sparse MoE (SMoE), activates only a small subset of subnetworks for each input [25, 26]. This approach preserves the expressive power of MoE while significantly reducing computational overhead during inference. A well-known implementation is the switch Transformer [26], where the FFN in a Transformer [27] is replaced with a set of experts of equal size, and only one expert is selected per input. Expert selection is managed by a router network $\mathcal{G}(\cdot)$, implemented as a fully connected layer parameterized with $\boldsymbol{W}_g$ and $\boldsymbol{b}_g$:

$$\boldsymbol{p}(\boldsymbol{x}) = \mathrm{softmax}(\boldsymbol{W}_g\boldsymbol{x} + \boldsymbol{b}_g), \quad \mathcal{G}(\boldsymbol{x}) = \text{top-1}(\boldsymbol{p}(\boldsymbol{x})) \tag{1}$$

where $\boldsymbol{x}$ is an input representation and $\text{top-1}(\cdot)$ is a selection function that outputs the largest value. The final output of the SMoE module is calculated as $\mathcal{G}(\boldsymbol{x})E_i(\boldsymbol{x})$, where i is the index of the selected expert and $E_i(\boldsymbol{x})$ is its output.

SMoE-based models with a trainable router often suffer from the routing imbalance issue that only a small subset of experts is frequently selected, while the others remain undertrained. To encourage balanced expert assignment, the following balancing loss function is commonly used [25, 28].

$$\mathcal{L}_{balance}(\boldsymbol{x}) = \sum_{j=1}^{n_{expert}} (\boldsymbol{p}_j(\boldsymbol{x}) - \frac{1}{n_{expert}}) \tag{2}$$

where n_{expert} is the total number of experts and $\boldsymbol{p}_j(\boldsymbol{x})$ is the j-th element of the selection probability vector $\boldsymbol{p}(\boldsymbol{x})$ in (1).

3. Proposed method

We propose an approach to enhance the model capacity to better accomodate various languages or language families (§3.2), and promote knowledge sharing among distant languages (§3.3).

3.1. Switch Conformer

Our framework is based on Conformer [5], in which we replace the second FFN module with an SMoE consisting of n_{expert} independent experts, forming what we term the *switch Conformer* block. This allows the model to effectively handle highly heterogeneous data. Each expert has the same structure as the Conformer FFN, consisting of two linear transformations with a ReLU activation function in between, defined as:

$$E_i(\boldsymbol{x}) = \boldsymbol{W}_2\,\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{3}$$

where $W_1 \in \mathbb{R}^{d_{model} \times d_{expert}}$ and $W_2 \in \mathbb{R}^{d_{expert} \times d_{model}}$. The expert FFN dimension d_{expert} is set equal to the baseline Conformer FFN dimension d_{ff}. As illustrated in Figure 1, a trainable router selects a single expert from the SMoE using the `top-1` selection function, following the switch Transformer paradigm.

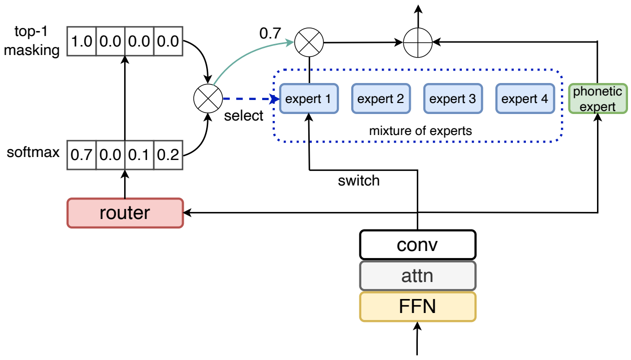

Figure 1: *Switch Conformer block with shared phonetic expert.*

This design choice is motivated by the following considerations. First, previous studies have shown that Conformer consistently outperforms the vanilla Transformer in various ASR tasks [29, 18]. Second, in our preliminary experiments, replacing the first FFN module with the SMoE led to significantly worse performance compared to replacing the second FFN. Third, increasing the number of active experts also negatively affected performance, possibly due to the reduced capacity per expert to maintain constant computational cost.

As illustrated in Figure 2(a), we train the entire network, consisting of L switch Conformer blocks, using the primary grapheme-based targets. To further mitigate the routing imbalance mentioned in §2.3, we introduce the *expert dropout* strategy in combination with the balancing loss. In this approach, each expert is randomly dropped with a probability of 0.1, preventing specific experts from being disproportionately selected. Because this is harmful in later training steps, we apply this strategy only for the first 5k steps.

3.2. Universal phonetic experts

To promote cross-lingual knowledge sharing, we introduce the universal phonetic expert along with the SMoE module in each block. It is implemented as an FFN with a capacity d_{shared}, determined by the capacity ratio c_{shared} as $d_{shared} = c_{shared} \cdot d_{ff}$. When incorporating this shared expert, the FFN dimension of an expert in SMoE is reduced to $d_{expert} = (1 - c_{shared}) \cdot d_{ff}$. Unlike SMoE, which selectively routes inputs, the universal phonetic expert processes all inputs. Thus, the output of the combined module is calculated as $\mathcal{G}(\boldsymbol{x})E_i(\boldsymbol{x}) + E_{shared}(\boldsymbol{x})$.

This phonetic expert is designed to capture universal phonetic features by training against IPA sequences as targets. As illustrated in Figure 2(b), during loss calculation with IPA targets $\boldsymbol{y}_{ipa}$, all SMoE experts are deterministically dropped by zeroing out all elements in the selection probability vector $\boldsymbol{p}$, so that the shared expert explicitly learns to specialize in acoustic-to-phoneme mapping. In contrast, when training with primary grapheme targets $\boldsymbol{y}_g$, both the SMoE and the phonetic expert

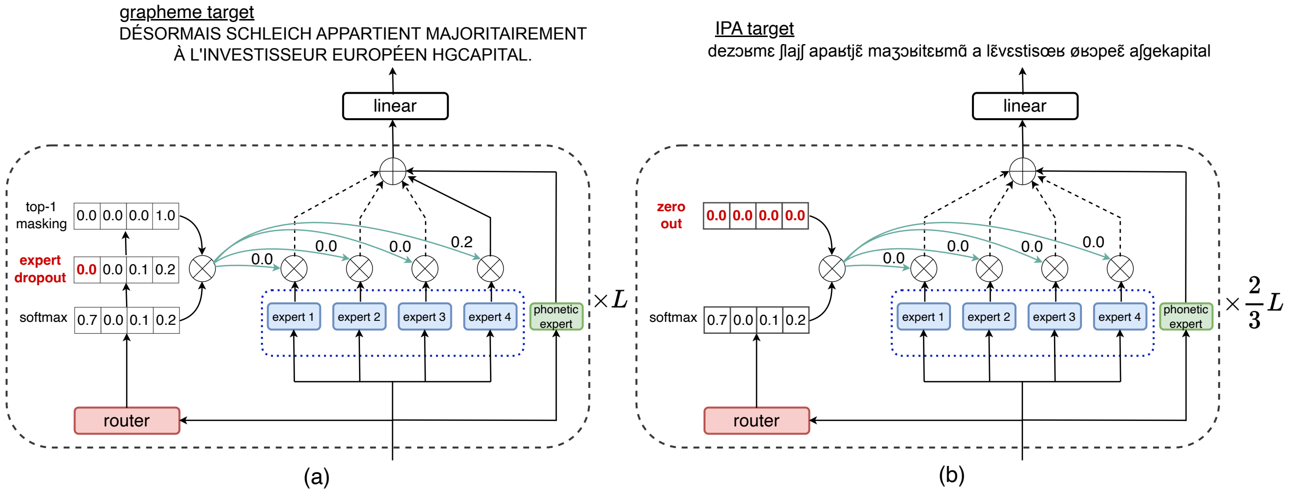

Figure 2: *(a) Expert dropout and (b) target-based routing strategies for training on grapheme and phonetic targets, respectively. Convolution, attention and FFN modules are omitted in the switch Conformer block for clarity.*

contribute to the encoder output (Figure 2(a)). We refer to this simple yet effective approach as *target-based routing* training for phonetic experts. The overall loss function is given by:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{rnnt}(\boldsymbol{y}_g, \boldsymbol{x}^L) + 0.3 \cdot \mathcal{L}_{ctc}(\boldsymbol{y}_g, \boldsymbol{x}^L) + \\ &0.1 \cdot \mathcal{L}_{ctc}(\boldsymbol{y}_{ipa}, \overline{\boldsymbol{x}}^{\frac{3}{4}L}) + 0.1 \cdot \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}_{balance}(\boldsymbol{x}^{l-1}) \end{aligned} \tag{4}$$

where $\boldsymbol{x}^L$ represents the output of the L-th encoder layer, and $\overline{\boldsymbol{x}}^{\frac{3}{4}L}$ is the output of the $\frac{3}{4}L$-th layer computed using the target-based routing. Following [17, 16], the IPA loss is applied at a lower layer, based on the observation that phonetic information is encoded at shallower layers to graphemes. $\mathcal{L}_{ctc}$ and $\mathcal{L}_{rnnt}$ are the loss functions of the CTC [1] and RNN-T [2] criteria.

4. Experimental evaluations

4.1. Datasets

For our experiments, we constructed two datasets from CommonVoice [30] v16.1. The first, *WE-5langs*, consists of five major Western European languages: German (`de`), English (`en`), Spanish (`es`), French (`fr`), and Italian (`it`). This set evaluates the effectiveness of our approach in high-resource, geographically proximate languages. The second, *Global-5langs*, includes a linguistically and geographically diverse selection: Arabic (`ar`), Bengali (`bn`), Russian (`ru`), Swahili (`sw`), and Thai (`th`). This set assesses model performance on low-resource languages with distinct writing systems and linguistic characteristics. To ensure statistical reliability, we selected languages with at least 30 hours of training data. We also created *All-10langs*, a combined set of *WE-5langs* and *Global-5langs*, representing a practical scenario where low-resource ASR training is augmented with high-resource language data. Table 1 summarizes the dataset specifications.

Because we found that no single g2p tool gives consistently accurate results across languages, we used three different tools for g2p conversion, namely, the ByT5-based g2p tool (Charsiu) [19] (`ar`, `de`, `en`, `es`, `fr`, `it`, `ru`, `sw`), Phonetisaurus [20] (`bn`) and Epitran [21] (`de`, `th`). We build vocabularies of 0.5k, 2k and 3k word pieces using byte pair encoding [31] for monolingual, 5-lang and 10-lang models, respectively. All multilingual models with the phonetic experts share the same IPA vocabulary consisting of 247 distinct symbols.

Table 1: *Specifications of 10 languages in CommonVoice v16.1*

code	subfamily	region	train (h)
`de`	Germanic	Western Europe	913.7
`en`	Germanic		1720.8
`es`	Romance,Italic		473.1
`fr`	Romance,Italic		777.1
`it`	Romance,Italic		247.9
`ar`	Afro-Asiatic	Middle East	32.3
`bn`	Indo-Iranian	South Asia	33.7
`ru`	Slavic	Eastern Europe	37.8
`sw`	Atlantic-Congo	Sub-Saharan Africa	69.4
`th`	Kra-Dai	South-East Asia	37.4

4.2. Models

In our experiments, we focus on streaming multilingual ASR based on RNN-Transducer [2], because improving computational efficiency is more critical in the streaming setting. We use 80-dimensional log-Mel filterbank outputs, extracted with a window size of 25ms at every 10ms, as the input features. The feature sequences are then subsampled using a 2-layer 2D convolutional network with 256 filters, a kernel size of 3 and a stride of 2, at a rate of 4, before fed into the encoder.

All the models consist of 12 chunk-wise Conformer blocks [32, 33], and the model dimension d_{model}, FFN dimension d_{ff} and the number of attention heads are set to be 512, 2048, and 8, respectively. We used the chunk size of 20 and the history size of 20 for streaming chunk processing. We use the RNN-T architecture [2] with a prediction network consisting of a one-layer unidirectional LSTM with 512 cells, and a feed-forward joint network with 640 cells. We take the IPA loss at the 8-th layer as shown in Figure 2 (b). We set the capacity ratio of the phonetic expert c_{shared} to be $\frac{1}{16}$, and thus $d_{expert} = 1920$ and $d_{shared} = 128$. We set the number of experts in the SMoE, n_{expert}, to be 8. We add these experts to all of 12 Conformer blocks when building the proposed models.

All models were trained using the Adam optimizer [34] with a linear-warmup of 25k steps and a peak learning rate of 0.0015. We evaluated the models in word error rate (WER) for *WE 5-langs*, while used character error rate (CER) for *All 10-*

Table 2: *Results for WE-5langs (WER / CER (%))*

training data	model	# active params	de	en	es	fr	it	ave.
monolingual	Conformer	83M	13.4 / 4.1	25.7 / 11.8	19.3 / 6.4	17.7 / 6.0	30.3 / 8.5	21.2 / 7.4
WE-5langs	Conformer	87M	14.0 / 3.9	24.2 / 10.7	13.1 / 4.0	19.0 / 6.3	15.9 / 4.1	17.3 / 5.7
	switch Conformer	87M	12.4 / 3.4	22.6 / 10.0	11.8 / 3.5	17.2 / 5.7	13.7 / **3.5**	15.6 / 5.1
	+ phonetic expert	87M	**12.2 / 3.3**	**22.1 / 9.7**	**11.5 / 3.4**	**16.9 / 5.5**	**13.6 / 3.5**	**15.3 / 5.0**

Table 3: *Results for All-10langs (CER (%))*

training data	model	# active params	de	en	es	fr	it	ave.	ar	bn	ru	sw	th	ave.
monolingual	Conformer	83M	4.1	11.8	6.4	6.0	8.5	7.4	51.6	82.4	15.1	11.8	19.8	34.6
All-10langs	Conformer	90M	4.0	10.9	4.1	6.5	4.2	5.8	35.9	11.2	12.7	9.5	14.4	14.6
	switch Conformer	90M	3.5	10.0	**3.5**	5.7	**3.6**	5.2	36.7	9.7	10.9	8.8	13.8	13.6
	+ phonetic expert	90M	**3.4**	**9.9**	**3.5**	**5.6**	**3.6**	**5.1**	**35.0**	**8.7**	**10.4**	**8.5**	**12.3**	**12.7**

langs, because some languages in this set do not have explicit word boundaries in their writing systems.

4.3. Results on WE-5langs

Table 2 presents the results for the monolingual and multilingual models evaluated on WE-5langs. Comparing the monolingual and multilingual Conformer baselines, we observe that multilingual training reduces WER except for de and fr, which already have large training datasets and relatively lower monolingual WERs. The switch Conformer consistently and significantly outperformed the baseline across all five languages. In particular, it achieved lower WERs than the monolingual models even for de and fr, indicating that the SMoE module effectively mitigates the negative transfer observed in the baseline Conformer. Adding phonetic experts led to further improvements across all five languages, despite that additional phonetic supervision has been shown to be less effective for languages sharing similar graphemes in a previous study [16]. This supports the effectiveness of our approach compared to conventional models that do not use a dedicated subnetwork specialized in predicting phonetic units from speech. Both improvements in WER from the use of the SMoE and the phonetic expert are statistically significant at the 1% level.

4.4. Results on All-10langs

Table 3 compares the results for models trained on *ALL-10langs*. Incorporating the SMoE module consistently improved performance over baseline monolingual and multilingual Conformer models, except for ar, where the baseline multilingual Conformer achieved a slightly lower CER. Notably, the Switch Conformer with universal phonetic experts achieved the lowest CERs across all languages, including ar.

Comparing the results for *WE-5langs* and *Global-5langs*, we find that using universal phonetic information is particularly effective for low-resource languages with distinct graphemes from the Western languages. The large gains in bn and th can be attributed to the relatively small numbers of IPA units used in these languages (51 and 42), which increased the number of training examples per unit. These results suggest that phonetic experts facilitate knowledge sharing across distant languages, while SMoE effectively captures language-specific variations. On average, the addition of the phonetic experts reduced CER by 0.9 points over the vanilla switch Conformer for *Global-5langs*, a statistically significant improvement at the 1% level.

Table 4: *Ablation on proposed techniques (CER (%)).*

	WE-5langs	other 5 langs
switch Conformer w/o expert drop	5.3	13.8
switch Conformer	5.2	13.6
+ shared expert	5.2	13.6
+ IPA auxiliary task	5.2	13.2
+ target-based routing	**5.1**	**12.7**

We conducted an ablation study to evaluate the impact of the techniques introduced in this work. As shown in Table 4, expert dropout training (rows 1–2) effectively enhances the performance of the switch Conformer. The third row (“+ shared expert ”) demonstrates that simply adding shared experts without assigning them an explicit role does not provide any gains while not harmful, likely due to the slightly reduced capacity per expert. The fourth row presents results obtained using an IPA auxiliary task without our target-based routing strategy. Comparing this with the final row, we see that separately routing inputs for IPA and grapheme targets, as discussed in §3.2, is crucial for achieving meaningful improvements.

5. Related work

Unlike their fundamental role in recent LLM advancements (e.g., [35][28]), sparse experts remain underexplored in ASR. SpeechMoE [36] and its successor, SpeechMoE2 [36], were the first to apply MoE to ASR, demonstrating its effectiveness in multi-domain and multi-accent ASR tasks using CTC-based offline models. Our approach differs in two key ways: we enhance capacity of the state-of-the-art Conformer for better performance, and more importantly, incorporate universal phonetic information to address issues inherent to multilingual ASR.

6. Conclusion

This paper proposed a novel architecture that enhances model capacity while effectively balancing language-universal knowledge and language-specific features to improve multilingual ASR. Our approach significantly improved performance in major Western European languages and proved particularly effective for low-resource languages. Future work will focus on incorporating more universal articulatory features for better adaptability to low-resource languages [22, 23]. We are also interested in layer-wise analysis of expert assignment by the trainable router for different languages, speakers and phonemes.

7. References

[1] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification : Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. of the 23st International Conference on Machine Learning*, 2006, pp. 369–376.

[2] A. Graves, “Sequence transduction with recurrent neural networks,” in *LCML*, 2012, pp. 4945–4949.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *ICASSP*, 2016, pp. 4945–4949.

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

[5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented Transformer for Speech Recognition,” in *Interspeech*, 2020, pp. 5036–5040.

[6] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, “Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration,” in *Interspeech*, 2019, pp. 1408–1412.

[7] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” in *Interspeech*, 2015, pp. 3586–3589.

[8] 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,” in *Interspeech*, 2019, pp. 2613–2617.

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

[10] 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,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, pp. 3451–3460.

[11] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, “Self-supervised learning with random-projection quantizer for speech recognition,” in *Proceedings of ICML*, 2022, pp. 3915–3924.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[13] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” in *arXiv preprint arXiv:2006.13979*, 2020.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” *arXiv preprint arXiv:2212.04356*, 2022.

[15] S. Watanabe, T. Hori, and J. R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in *ASRU*, 2017, pp. 265–271.

[16] S. W. Oliver Adams, Matthew Wiesner and D. Yarowsky, “Massively Multilingual Adversarial Speech Recognition,” in *NAACL*, pp. 96—-108.

[17] J. Lee, M. Mimura, and T. Kawahara, “Leveraging IPA and articulatory features as effective inductive biases for multilingual asr training,” in *ICASSP*, 2025.

[18] A. Gulati *et al.*, “Conformer: Convolution-augmented Transformer for Speech Recognition,” in *Interspeech*, 2020, pp. 5036–5040.

[19] J. Zhu1, C. Zhang, and D. Jurgens, “Byt5 model for massively multilingual grapheme-to-phoneme conversion,” in *Interspeech*, pp. 446–450.

[20] J. R. Novak, N. Minematsu, and K. Hirose, “Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework,” in *Natural Language Engineering*, vol. 22, 2016, pp. 907–938.

[21] D. R. Mortensen, S. Dalmia, and P. Littell, “Epitran: Precision G2P for Many Languages,” in *LREC*, 2018.

[22] S. Stuker, T. Schultz, F. Metze, and A. Waibel, “Multilingual articulatory features,” in *ICASSP*, 2003.

[23] J. Lee, M. Mimura, and T. Kawahara, “Embedding articulatory constraints for low-resource speech recognition based on large pre-trained model,” in *Interspeech*, pp. 1394—-1398.

[24] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” in *Neural Computation*, 1991, pp. 79–87.

[25] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” in *ICLR*, 2017.

[26] W. Fedus, B. Zoph, and N. Shazeer, “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity,” in *Journal of Machine Learning Research*, vol. 23, pp. 5232–5270.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. ukasz Kaiser, and I. Polosukhin, “Attention is all you need,” in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, vol. 30, 2017.

[28] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, “GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding,” in *ICLR*, 2021.

[29] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, “Recent Developments on Espnet Toolkit Boosted By Conformer,” in *ICASSP*, pp. 5874–5878.

[30] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common Voice: A Massively-Multilingual Speech Corpus,” in *LREC*, vol. 23, pp. 4218—-4222.

[31] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2016, pp. 1715–1725.

[32] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, “Developing real-time streaming transformer transducer for speech recognition on large-scale dataset,” in *ICASSP*, 2021, pp. 5904–5908.

[33] M. Mimura, T. Moriya, and K. Matsuura, “Advancing Streaming ASR with Chunk-wise Attention and Trans-chunk Selective State Spaces,” in *ICASSP*, 2025.

[34] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.

[35] A. Q. Jiang *et al.*, “Mixtral of Experts,” in *arXiv:2401.04088*, 2024.

[36] Z. You, S. Feng, D. Su, and D. Yu, “SpeechMoE: Scaling to large acoustic models with dynamic routing mixture of experts,” in *Interspeech*, pp. 2077—-2081.