AMuSE: Attentive Multilingual Speech Encoding for Zero-Prior ASR

Ashutosh Varshney, Debmalya Chakrabarty, Akshat Jaiswal, Harish Arsikere, Abhinav Jain, Swayambhu Nath Ray, Frederick Weber, Anand Mohan, Prantik Sen, Garima Lalwani, Sambuddha Bhattacharya, Sri Garimella
Amazon AGI, Bangalore
{varshnas, debmac, akshatkj, arsikere, jabhinv, swayar, fweber, anamhan, sprantik, glalwani, bsambudd, srigar}@amazon.com

***Abstract*—Multilingual ASR offers training, deployment and overall performance benefits, but models trained via simple data pooling are known to suffer from cross-lingual interference. Oracle language information (exact-prior) and language-specific parameters are usually leveraged to overcome this, but such approaches cannot enable seamless, truly multilingual experiences. Existing methods try to overcome this limitation by relying on inferred language information or language agnostic mixture-of-experts, but they incur additional runtime complexity and/or training cost in addition to being less effective in streaming scenarios. Building on previous studies where models were trained to handle mixed-prior (knowledge that the underlying language belongs to a known group), we propose Attentive Multilingual Speech Encoding (AMuSE), a training framework designed to match exact-prior performance even in the absence of underlying language information at runtime (zero-prior), thereby making the model prior-agnostic. Leveraging AMuSE, we build a zero-prior enabled LLM-based ASR system that outperforms several exact-prior driven state-of-the-art benchmarks.**

***Index Terms*—Multilingual ASR, Speech encoder, LLM**

I. Introduction

Multilingual ASR is an active area of research due to its general accuracy benefits and potential for simplified model training and deployment across languages. Methods leveraging oracle language information (exact-prior) and language-specific parameters [1], [2] achieve good performance but are not truly multilingual, as they cannot operate when no language information is known apriori (zero-prior). Common industry approaches to provide a true multilingual experience to the user include techniques such as explicit language identification (LID) followed by decoding, joint ASR-LID training, and parallel monolingual decoding runs with LID arbitration [3]–[7], but such methods incur additional runtime latency and cost.

Mixture-of-experts (MoE) based approaches have gained popularity for zero-prior multilingual ASR, wherein a fixed number of experts are dynamically activated during inference [8]–[11]. However, the routing strategies employed for activating the experts are purely data-driven, potentially leading to sub-optimal performance, particularly if the routing network does not generalize well. These approaches are also not designed to take advantage of any language information that might be available at runtime. Moreover, MoEs are less effective for streaming use cases [9], where some degree of implicit language awareness is crucial. Mixture of informed experts (MIE) instead exploits the language information available in data during training, by pre-assigning each expert to a particular language group [12]. Alternatively, configurable multilingual model (CMM) [13] trains language experts via a mixed-conditioning strategy, i.e., randomly activating several experts in addition to oracle expert for simulating runtime user selection. This approach allows the model to operate when it is known that the underlying language belongs to a given subset during inference (mixed-prior). However, CMM is constrained by uniform weighting of experts for aggregation, making it challenging to maintain a reasonable multilingual ASR performance as the number of languages increases during inference.

To address the aforementioned shortcomings of existing works, we introduce a training framework called **A**ttentive **Mu**ltilingual **S**peech **E**ncoding (**AMuSE**). Speech encoders trained using AMuSE enable downstream ASR systems to operate seamlessly in any mode (exact-prior, mixed-prior, or zero-prior) based on user selection, with almost similar performance. This is accomplished through a novel integration of an attention mechanism over language experts, coupled with a masking strategy that uses mixed-conditioning of language priors. In contrast to CMM, weights for combining experts in an AMuSE trained encoder are dynamically computed based on the acoustics of input speech via an attention mechanism. Furthermore, unlike other MoE-based approaches, language information available in training data is utilized to train experts via a mixed-conditioning based masking strategy. AMuSE thus offers the best of MoEs and CMM to provide models that are agnostic to prior language information. Results show that our proposed framework is effective for both streaming and non-streaming use cases, with minimal gap between exact-prior and zero-prior ASR performance. In benchmarking against several state-of-the-art exact-prior driven ASR systems, we demonstrate that AMuSE can be leveraged to outperform them in zero-prior mode by interfacing with a pre-trained text LLM.

II. Methodology

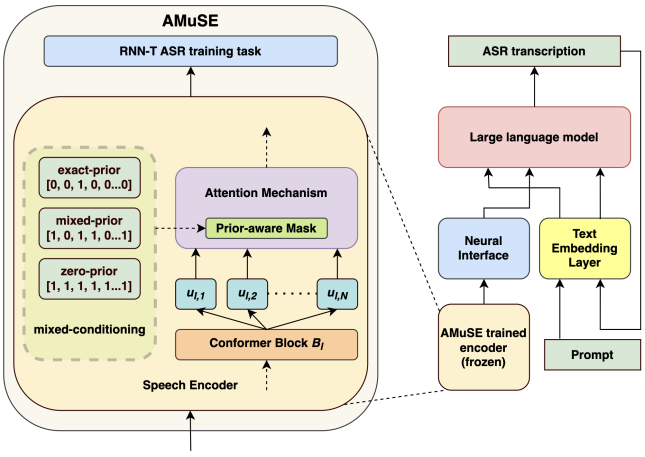

Fig. 1. AMuSE framework for training speech encoder (left). Interfacing of the obtained encoder with a pre-trained LLM (right).

AMuSE comprises of three key components to make a speech encoder ASR-aware and agnostic to prior language information: (i) attention over language experts to dynamically compute their individual weights for aggregation, (ii) masking in attention computation guided by mixed-conditioning of language priors, and, (iii) training via RNN-

T based ASR task [14]. To further obtain a zero-prior enabled LLM-based ASR system, AMuSE trained encoder is frozen and interfaced with a pre-trained text LLM for speech-text LLM training [15]. All these details are depicted in Figure 1.

A. Encoder architecture and training strategy

We use the Conformer architecture [16] along with lightweight residual adapters [17] as language experts for the speech encoder. As shown in Figure 1, language experts $U_l = [u_{l,1}, u_{l,2}, \ldots, u_{l,N}]$ are placed after the l^{th} Conformer block B_l ($l \in [1, L]$), where N and L represent the number of languages and Conformer blocks, respectively. To keep the encoder size in check as N grows and reduce memory consumption, experts are only placed after the last 4 Conformer blocks. Intermediate output from block B_l is passed through the language experts U_l and aggregated into a single representation via a prior-aware attention mechanism (Section II-B). To facilitate integration into a specific downstream task, the encoder should be equipped with task-specific awareness. The encoder in AMuSE is therefore trained with an RNN-T task to make it ASR-aware.

Our framework is motivated by the mixed-conditioning approach proposed in CMM [13], which uses language experts to support mixed-priors during inference. The approach randomly selects a group of secondary language experts, in addition to the ground-truth language expert, for training on each input sample. Individual outputs from these experts are uniformly weighted for aggregation, which is, however, bottlenecked by the maximum number of languages to be supported at runtime. This limitation arises because the contribution from the expert of ground-truth language gets diluted as the number of languages increases, at worst to $1/N$ in zero-prior mode. The idea of mixed-conditioning of language priors is developed further in AMuSE to address this limitation of CMM and improve model performance in mixed-prior and zero-prior settings.

We propose that to reduce the model's reliance on language information while maintaining performance across prior conditions (from exact-prior to zero-prior), experts should be aggregated with adaptive weights w^* that adjust to the acoustics of input speech and handle variations in priors during inference. Building upon the mixed-conditioning strategy from CMM, we extend it to simulate runtime prior scenarios more effectively during training. For each training sample, the prior mode is first probabilistically sampled from: (i) exact-prior, which entails knowing the user language apriori; (ii) mixed-prior, which assumes that the underlying language belongs to a known subset of languages; and (iii) zero-prior, which is language agnostic. This sampling ensures that the training is balanced by avoiding potential over-reliance on either exact-prior or zero-prior cases. Depending on this selection, a number of secondary languages (0 for exact-prior, N-1 for zero-prior) are subsequently chosen in addition to the ground-truth language, similar to CMM. An attention mechanism is then introduced over the language experts and is guided by this selection of languages to obtain the weights w^*, as explained in the following section.

B. Attention mechanism with prior-aware masking

Our motivation is to improve upon the uniformly weighted aggregation of outputs from experts, as it dilutes the contribution of an individual language within a group of multiple languages. We allow experts to be trained with a pool of multilingual data and learn their individual contributions on the fly using a prior-aware attention mechanism to achieve this. Extending the previous works on obtaining context vector for attention computation using vector concatenation followed by down-projection to a single value [18]–[20], we additionally apply a prior-aware mask (mask_{PA}) that utilizes the mixed-conditioning of language priors explained in Section II-A. This approach facilitates the learning of contribution from individual experts such that the aggregated representation encapsulates both language-specific (learning expert through exact-prior) as well as language-agnostic nuances (aligning the expert to mixed-prior and zero-prior modes). More formally, given x_l as the output of the l^{th} Conformer block B_l, adaptive weights w^* to aggregate representations from language experts are obtained as follows (Note that the equations are indexed with $i \in \{1, \ldots, N\}$ and vectorised across the N languages):

$$
\begin{aligned}
o_l[i] &= u_{l,i}(x_l) && \{o_l \in \mathbb{R}^{N \times d_{model}}\} \\
h_l[i] &= \tanh([o_l[i]; x_l] \cdot w_{l,1,i}) && \{h_l \in \mathbb{R}^{N \times d_{hidden}}\} \\
\widetilde{h_l}[i] &= \text{mask}_{\text{PA}}(h_l[i] \cdot w_{l,2,i}) && \{\widetilde{h_l} \in \mathbb{R}^{N \times 1}\} \\
w^* &= \text{softmax}(\widetilde{h_l}) && \{w^* \in \mathbb{R}^{N \times 1}\}
\end{aligned}
$$

$W_{l,1} = [w_{l,1,i} \in \mathbb{R}^{2*d_{model} \times d_{hidden}}]$ and $W_{l,2} = [w_{l,2,i} \in \mathbb{R}^{d_{hidden} \times 1}]$ $\forall i \in \{1, \ldots, N\}$ are the trainable parameters. d_{model} and d_{hidden} are encoder hyper-parameters detailed in Table I. o_l vector represents the language specific outputs from each of the N experts in U_l. The context vector h_l is obtained by concatenating outputs $o_l[i]$ with x_l (results in dimension $\mathbb{R}^{2*d_{model}}$), projecting them with $W_{l,1}$ and applying tanh non-linearity. $\widetilde{h_l}$ is then computed by down-projecting h_l with $W_{l,2}$ and applying mask_{PA}. mask_{PA} ensures the simulation of language prior scenarios by setting $\widetilde{h_l}[i] = -\infty$ if the i^{th} expert $u_{l,i}$ was not chosen by the mixed-conditioning selection for current training sample. The same masking logic is also used during inference according to the selected prior. w^* is finally obtained by applying softmax on $\widetilde{h_l}$. These computed weights are used to combine outputs o_l from language experts as follows:

$$
x_{l+1} = B_{l+1}(\sum_{i=1}^{N}(o_l[i] * w^*[i])) \qquad \{x_{l+1} \in \mathbb{R}^{d_{model}}\}
$$

This prior-aware attention mechanism, guided by the mixed-conditioning of language priors, improves the downstream model's performance across all possible prior modes (exact-prior, mixed-prior, and zero-prior).

TABLE I
AMuSE encoder hyper-parameters

	AMuSE-small	*AMuSE-large*
d_{model}	800	2048
d_{ff}	1600	4096
d_{hidden}	128	128
L	18	17
N	5	5
causal	True	False
attention heads	8	8
head size	64	128
left context	60	-
kernel size	32	32
output size	1280	1280
total parameters	200 million	1 billion
supported languages	{en, fr, de, it, es}	{en, fr, de, it, es}

TABLE II

AMuSE-small WER results on Vox Populi, evaluated across prior modes. (i) exact-prior: Only the oracle language expert is active. (ii) mixed-prior: Multiple secondary language experts are active in addition to oracle expert. Secondary languages are randomly sampled. (iii) zero-prior: All the experts are active.

Language	exact-prior		mixed-prior (2 languages)		mixed-prior (3 languages)		zero-prior		
	Baseline	*AMuSE-s*	*Baseline*	*AMuSE-s*	*Baseline*	*AMuSE-s*	*Baseline*	*AttentionMoE*	*AMuSE-s*
English	12.8	12.5	13.0	12.8	13.1	12.8	13.1	13.3	12.9
French	16.5	16.1	16.7	16.4	17.2	16.7	18.6	17.0	17.0
German	16.4	16.2	16.5	16.3	16.9	16.4	18.2	17.1	16.6
Italian	28.6	28.3	28.9	28.5	29.2	28.6	31.9	30.5	29.9
Spanish	12.9	12.9	13.1	12.9	13.3	12.9	14.1	14.1	13.2
Average	17.4	**17.2**	17.6	**17.4**	17.9	**17.5**	19.2	18.4	**17.9**

III. Experiments and Results

A. ASR-aware streaming encoder via AMuSE

To demonstrate the efficacy of our proposed method on ASR performance across different language prior modes, we develop an ASR-aware streaming encoder called *AMuSE-small(s)*, with hyper-parameters detailed in Table I. The RNN-T task used for training within the AMuSE framework consists of a 4-layer LSTM [21] network with 1280 units as the decoder, followed by a feed-forward layer as the joint network. *AMuSE-s* is trained on a mix of 3.5M hours of anonymized internal data and 100K hours of publicly available corpora for 500K steps until convergence. For comparison, we train a *Baseline* encoder with the same hyper-parameters and training setup as *AMuSE-s*, but by assigning uniform weights to language experts that are selected via mixed-conditioning as proposed in CMM (i.e., the attention mechanism is removed).

Both the encoders are evaluated on the Vox Populi test set [22] across different runtime language prior conditions using the RNN-T setup. Table II demonstrates that the word error rate (WER) of *AMuSE-s* improves relative to *Baseline* with increase in number of languages at runtime. Our encoder is marginally better than *Baseline* in the exact-prior mode, where the ground-truth language is known. As we introduce some degree of cross-lingual interference in mixed-prior mode by presenting more languages, results indicate that average WER of *AMuSE-s* remains consistent with increase in number of languages, in contrast to some degradation in *Baseline* performance. The advantage of our framework is most evident in zero-prior mode which has maximum multilingual confusion; *AMuSE-s* achieves 6.8% better relative WER compared to *Baseline* on average. Moreover, the relative gap between exact-prior and zero-prior results for *AMuSE-s* is only 4.1%, compared to 10.3% for *Baseline*.

TABLE III

AttentionMoE WER results on Vox Populi across different prior modes highlight the significance of utilizing language prior information in AMuSE.

Language	exact-prior	mixed-prior (2 languages)	zero-prior
English	20.0	18.6	13.3
French	97.0	73.8	17.0
German	91.9	90.9	17.1
Italian	97.3	91.8	30.5
Spanish	61.0	21.7	14.1
Average	73.4	59.4	18.4

An important detail of AMuSE is that language prior information is explicitly provided to the attention mechanism using mixed-conditioning strategy. This is achieved by a prior-aware masking function mask_{PA} described in Section II-B. To elucidate its importance, we train an encoder called *AttentionMoE* (identical to *AMuSE-s*) by removing mask_{PA} in w^* computation, i.e., employing only zero-prior training with pure attention. WER comparison between *AMuSE-s* and *AttentionMoE* in zero-prior mode is reported in Table II for the Vox Populi test set. Results empirically demonstrate that lack of any language prior information while training leads to inferior ASR performance which can be attributed to cross-lingual confusion. This absence of prior knowledge during training also results in significant degradation in exact-prior and mixed-prior performance of *AttentionMoE* as shown in Table III, rendering it unusable when some language information is available at disposal.

Finally, we present a visualization of the weights w^* in Figure 2 when *AMuSE-s* is operated in zero-prior mode. The diagonal pattern in the plotted heatmap indicates that w^* values are skewed towards the ground-truth language, thereby highlighting the implicit language-awareness of AMuSE trained encoder in zero-prior mode. These set of results collectively demonstrate the capability of AMuSE in obtaining language-informed representations for developing downstream ASR systems that are agnostic to language priors at runtime, while also showcasing its effectiveness for streaming use cases.

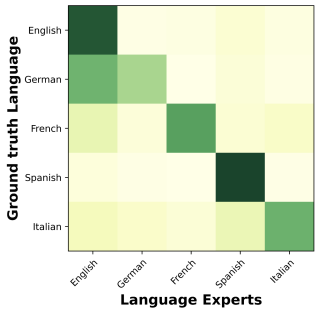

Fig. 2. Weights w^* used for combining language experts are skewed towards the oracle language for *AMuSE-small* in zero-prior mode during inference.

B. AMuSE trained encoder for LLM interfacing

In this section, we evaluate the synergy between zero-prior representations from an AMuSE trained speech encoder and a pre-trained text LLM serving as a decoder for ASR, by utilizing AcLLM framework [23]. The LLM is tasked with transcript generation for input speech by leveraging the AMuSE trained encoder alongside a task instruction prompt. A neural interface projects the speech representations from the encoder into the LLM's semantic space, as illustrated in Figure 1. The speech encoder is kept frozen, and only the LLM and neural interface are trained using cross-entropy loss on the output ASR transcription. We use an in-house multilingual

LLM based on LLaMA [24] with 1B parameters, pre-trained on 470B text tokens, and interface it with *AMuSE-s* for experimentation. For comparison, we also interface this LLM with a standard data-pooled Conformer encoder (referred as *Pooled*) that has identical hyper-parameters and RNN-T training task as *AMuSE-s* (Section III-A), trained to convergence.

We train both encoder-LLM models in two distinct settings: (i) without language information in the LLM prompt (e.g., "Transcribe what the speaker is saying") and (ii) with language information in the LLM prompt (e.g., "Transcribe what the Spanish speaker is saying"). All experiments were trained identically for 60K steps using approximately 50K hours of anonymized internal and external training data. Table IV reports the WER for these experimental models, averaged over individual results on Vox Populi and Multilingual Librispeech [25] test sets. Results clearly indicate that *AMuSE-s* outperforms the *Pooled* encoder when language information is not provided in the LLM prompt. Conversely, when the LLM is prompted with language information, WER degrades with the *Pooled* encoder, suggesting that the task of leveraging prior knowledge falls heavily on the LLM due to sub-optimal multilingual representations. In contrast, LLM language prompting is complementary to representations from *AMuSE-s*, improving ASR performance by 7.1% relative (15.5 to 14.4). These results empirically show that speech representations from an encoder trained with AMuSE are more effective for building an LLM-based multilingual ASR system due to their demonstrated synergy with language prompting.

TABLE IV
Comparison of *AMuSE-small* and *Pooled* encoders for ASR when interfaced with a pre-trained text LLM. An average of individual WER on Vox Populi and Multilingual Librispeech test sets is reported for each language.

	Pooled		*AMuSE-s*	
Language prompting	No	Yes	No	Yes
English	13.1	13.5	11.8	**11.2**
French	16.1	16.3	14.5	**14.2**
German	17.2	29.0	17.1	**14.2**
Italian	23.3	21.5	22.6	**21.9**
Spanish	11.4	11.3	11.4	**10.4**
Average	16.2	18.3	15.5	**14.4**

C. Zero-prior non-streaming ASR via AMuSE

We develop an LLM-based ASR system agnostic to prior language information by scaling up the experimental setup described in Section III-B. We first train a non-streaming ASR-aware speech encoder called *AMuSE-large(l)* (hyper-parameters detailed in Table I) using our proposed framework, focusing on (i) increased encoder parameters, (ii) full context representations, and, (iii) data volume and diversity. It is trained with the same RNN-T task as *AMuSE-s* (Section III-A) for 1.25M steps until convergence, using a mix of human-transcribed and machine-labeled data. The training corpus consists of 3.5M hours of anonymized internal data and 500K hours from publicly available sources, including Vox Populi, Multilingual Librispeech, CommonVoice [26], PeopleSpeech [27], FLEURS [28] and Librivox [29]. The frozen *AMuSE-l* encoder is subsequently interfaced in zero-prior mode with the pre-trained text LLM described in Section III-B, and the model is trained for 1M steps until convergence. The data volume used for this speech-text LLM training was also increased to 5.5M hours, primarily by adding high quality machine-labeled data.

The prior-agnostic and ASR-aware *AMuSE-l* encoder, when combined with the generative capabilities of LLM, enables us to outperform several exact-prior driven state-of-the-art systems. We benchmark our *AMuSE-l* + LLM system against OpenAI Whisper Large-v3 [30], AWS Transcribe [31], AssemblyAI Universal-1 [32] and Google Gemini 1.5-Pro [33] on four standard test suites: FLEURS, Vox Populi, Multilingual Librispeech, and CommonVoice. As illustrated in Figure 3, *AMuSE-l* + LLM outperforms all models with significant WER improvements when oracle language information is provided in the LLM prompt (exact-prior mode). Moreover, the WER gap between exact-prior and zero-prior modes (i.e., when language information is not prompted) for our model is within 0.2% absolute across test suites. This is attributed to the efficacy of our novel training framework for non-streaming use cases, in addition to the extensive parameter space of *AMuSE-l*. Hence, *AMuSE-l* + LLM system establishes a new performance benchmark, demonstrating robust multilingual ASR capabilities with consistent performance irrespective of prior language information.

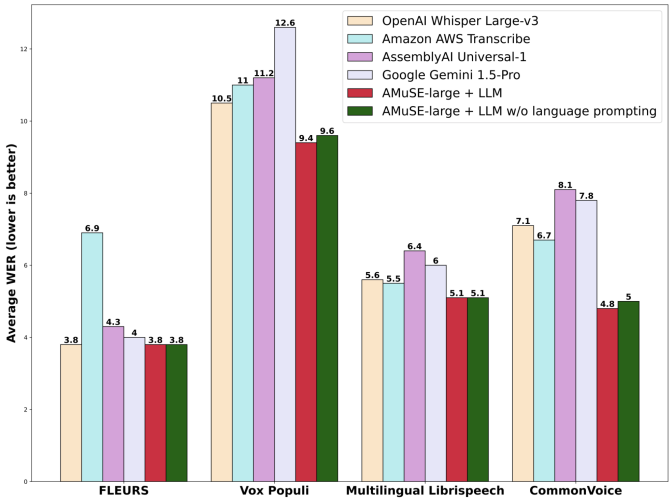

Fig. 3. Performance comparison of our *AMuSE-large* + LLM system against various state-of-the-art benchmarks. Reported WER for each test suite is an average over 5 languages (English, French, German, Italian, Spanish).

IV. Conclusion

This work introduces Attentive Multilingual Speech Encoding (AMuSE), a training framework designed to build ASR systems that are agnostic to prior language information, for both streaming and non-streaming use cases. Our results show that speech encoders trained via AMuSE are implicitly language-informed, with minimal gap between exact-prior and zero-prior ASR performance. We leverage AMuSE representations to develop an LLM-based ASR system that outperforms several state-of-the-art benchmarks, particularly in the challenging zero-prior setting. Consequently, AMuSE emerges as a robust framework, paving the way for future research avenues towards the realisation of truly multilingual ASR, enabling more satisfying and adaptable conversational experiences.

V. Acknowledgements

We thank Jahn Heymann, Ankish Bansal, Harish Mallidi, Phani Nidadavolu, Milind Rao, Nikhil Bhave, Bharat Padi, Venkata Kishore Nandury, Tuan Dinh, Andreas Schwarz and Vijay Girish for their valuable contribution to data preparation efforts.

REFERENCES

[1] J. Bai, B. Li, Q. Li, T. N. Sainath, and T. Strohman, "Efficient adapter finetuning for tail languages in streaming multilingual asr," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 841–10 845.

[2] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[3] C. Zhang, B. Li, T. Sainath, T. Strohman, S. Mavandadi, S.-Y. Chang, and P. Haghani, "Streaming end-to-end multilingual speech recognition with joint language identification," in *Interspeech 2022*, 2022, pp. 3223–3227.

[4] A. Waters, N. Gaur, P. Haghani, P. Moreno, and Z. Qu, "Leveraging language id in multilingual end-to-end speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 928–935.

[5] S. Wang, L. Wan, Y. Yu, and I. L. Moreno, "Signal combination for language identification," 2019. [Online]. Available: https://arxiv.org/abs/1910.09687

[6] C. Chandak, Z. Raeesy, A. Rastrow, Y. Liu, X. Huang, S. Wang, D. K. Joo, and R. Maas, "Streaming language identification using combination of acoustic representations and asr hypotheses," 2020. [Online]. Available: https://arxiv.org/abs/2006.00703

[7] S. Punjabi, H. Arsikere, Z. Raeesy, C. Chandak, N. Bhave, A. Bansal, M. Müller, S. Murillo, A. Rastrow, S. Garimella, R. Maas, M. Hans, A. Mouchtaris, and S. Kunzmann, "Streaming end-to-end bilingual asr systems with joint language identification," 2020. [Online]. Available: https://arxiv.org/abs/2007.03900

[8] W. Wang, G. Ma, Y. Li, and B. Du, "Language-routing mixture of experts for multilingual and code-switching speech recognition," in *Interspeech 2023*, 2023, pp. 1389–1393.

[9] K. Hu, B. Li, T. Sainath, Y. Zhang, and F. Beaufays, "Mixture-of-expert conformer for streaming multilingual asr," in *Interspeech 2023*, 2023, pp. 3327–3331.

[10] Y. Kwon and S.-W. Chung, "Mole : Mixture of language experts for multi-lingual automatic speech recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[11] Z. You, S. Feng, D. Su, and D. Yu, "Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts," in *Interspeech 2021*, 2021, pp. 2077–2081.

[12] N. Gaur, B. Farris, P. Haghani, I. Leal, P. J. Moreno, M. Prasad, B. Ramabhadran, and Y. Zhu, "Mixture of informed experts for multilingual speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6234–6238.

[13] L. Zhou, J. Li, E. Sun, and S. Liu, "A configurable multilingual model is all you need to recognize all languages," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6422–6426.

[14] A. Graves, "Sequence transduction with recurrent neural networks," 2012. [Online]. Available: https://arxiv.org/abs/1211.3711

[15] Z. Min and J. Wang, "Exploring the integration of large language models into automatic speech recognition systems: An empirical study," in *International Conference on Neural Information Processing*. Springer, 2023, pp. 69–84.

[16] 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*, 2020, pp. 5036–5040.

[17] K. Tomanek, V. Zayats, D. Padfield, K. Vaillancourt, and F. Biadsy, "Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021, pp. 6751–6760.

[18] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sep. 2015, pp. 1412–1421.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[20] S. Mitra, S. N. Ray, B. Padi, A. Sen, R. Bilgi, H. Arsikere, S. Ghosh, A. Srinivasamurthy, and S. Garimella, "Unified modeling of multi-domain multi-device asr systems," in *International Conference on Text, Speech, and Dialogue*, 2023, pp. 283–292.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 993–1003.

[23] Y. Bai *et al.*, "Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition," 2024. [Online]. Available: https://arxiv.org/abs/2407.04675

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

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

[26] 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 *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, May 2020, pp. 4218–4222.

[27] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021.

[28] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 798–805.

[29] LibriVox, "free public domain audiobooks," 2024. [Online]. Available: https://librivox.org/

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518.

[31] Amazon Web Services, "AWS Transcribe," 2024. [Online]. Available: https://aws.amazon.com/transcribe/

[32] AssemblyAI, "Universal-1: Robust and accurate multilingual speech-to-text," 2024. [Online]. Available: https://www.assemblyai.com/research/universal-1

[33] G. Team *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024. [Online]. Available: https://arxiv.org/abs/2403.05530