# Transfer Learning for Low-Resource Automatic Speech Recognition: A Survey and Future Directions

**Donghong Qin, Zhuo Chen, Xiao Liang*, Fengbo Bai**

AI and Big Data International Cooperation Joint Laboratory, School of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China
University Engineering Research Center of Computational Linguistics and Intelligence, Nanning 530006, China

**Keywords:** Speech recognition; Transfer learning; Multilingual learning; Self-supervised learning; Multi-task learning; Large Language model; Meta-learning

Citation: Qin D.H., Chen Z., Liang X., et al.: Transfer Learning for Low-Resource Automatic Speech Recognition: A Survey and Future Directions. Data Intelligence, 2026, Vol. 8, Art. No.: 20260082. DOI: https://doi.org/10.3724/2096-7004.di.2026.0082

## ABSTRACT

Most of the world's languages remain low-resource for automatic speech recognition (ASR). The bottleneck is not only the scarcity of labeled speech. It also includes strong variation in pronunciation, prosody, dialect, and domain, as well as the lack of linguistic tools and infrastructure. This survey reviews low-resource ASR through a unified transfer-learning perspective. We organize the literature into four connected strands: pretrain–adapt pipelines, parameter-efficient and domain-aware fine-tuning, task-based transfer through meta-learning and multi-task learning, and corpus expansion through augmentation and multimodal supervision. To compare methods that are usually reported on different tasks and metrics, we further summarize a normalized cross-study synthesis and relate it to a unified operational risk analysis based on effective sample size $m_{\text{eff}}$, structural compatibility $\mathcal{C}_{\text{eff}}$, and domain shift $\Gamma$. Beyond generic scaling trends, we pay particular attention to language-specific structure that is often under-modeled in low-resource ASR, especially tonal contrasts, tone sandhi, and rich or agglutinative morphology. We show how tone-aware constraints, F0-conditioned representations, morphology-aware output spaces, and auxiliary linguistic losses can complement self-supervised learning, parameter-efficient fine-tuning, and large multilingual speech models rather than replace them. The survey concludes with a synthesis roadmap that links large-scale speech foundation models, task-based transfer, language-specific inductive bias, and deployable adaptation, and with a set of concrete research questions for the next stage of low-resource ASR.

*Corresponding author: Xiao Liang (E-mail: 503377302@qq.com; ORCID: 0009-0006-5321-4565).

© 2026 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

## 1. INTRODUCTION

Automatic speech recognition has undergone several major technical transitions. Early systems relied on statistical modeling. They were followed by HMM-GMM pipelines, then by deep neural networks, and later by Transformer-based end-to-end models. This progression substantially improved recognition accuracy and broadened the application range of ASR. In resource-rich languages, especially English and Chinese, modern end-to-end systems already support mature industrial deployments.

This progress, however, is uneven. Most of the world's languages remain low-resource, as shown in Figure 1. For these languages, the difficulty is not only that speech corpora are small. Many of them also present linguistic properties that are poorly handled by English-centered modeling assumptions, including strong dialect variation, lexical tone, tone sandhi, rich inflection, and agglutinative word formation. These properties interact with data scarcity and make transfer more fragile.

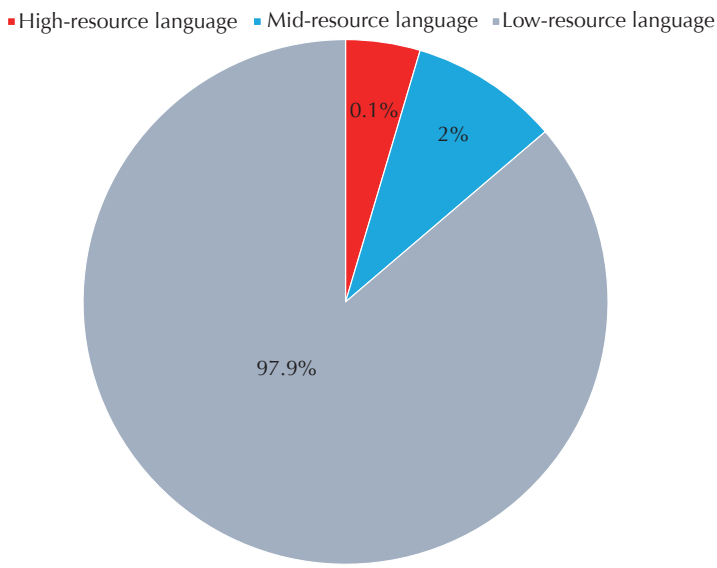



**Figure 1.** Proportion of High, Medium and Low Resource Languages for ASR.

The literature repeatedly shows that low-resource ASR is constrained by several factors at once. Linguistic and phonological complexity, including grammatical and semantic structure as well as accent variation, increases recognition difficulty [1-6]. At the same time, the available corpora are often too small, hard to collect, and insufficiently diverse [7-13]. Technical infrastructure is also limited in many settings [14-15]. As a result, methods that work well with large annotated corpora often degrade sharply in truly low-resource regimes.

Transfer learning has therefore become the central strategy for low-resource ASR. Empirical studies consistently show that conventional ASR systems trained from scratch perform poorly when labeled data are scarce, whereas pretraining followed by adaptation can deliver large gains [16-17]. Yet transfer learning

should not be understood as a generic recipe of "more pretraining" or "larger models." Different transfer mechanisms solve different bottlenecks. Some mainly build reusable representations. Some mainly convert unlabeled speech into supervision. Others mainly regulate how much of a pretrained model should move during adaptation. A useful survey should therefore explain not only what methods exist, but also how they relate to one another, where they saturate, and why some of them need language-specific structure to work reliably.

Two points motivate the present review. First, low-resource ASR is not only a small-data problem. It is also a structure-mismatch problem. If tone, sandhi, or morphology are left implicit, the model must infer them from extremely sparse supervision. Second, the strongest systems increasingly combine several transfer modes at once. A modern pipeline may begin with multilingual self-supervised learning, continue with parameter-efficient fine-tuning, exploit unlabeled target speech through self-training, and then rely on a large text model or a retrieval module at decoding time. Understanding such pipelines requires a common analytical language.

This survey has four aims. First, it organizes the field with a two-dimensional transfer-learning taxonomy. Second, it develops a unified operational analysis based on effective sample size, structural compatibility, and domain shift. Third, it pays special attention to language-specific transfer mechanisms for tonal and morphologically rich languages, which are often acknowledged but not analyzed in sufficient depth. Fourth, it closes with a synthesis roadmap and a small set of concrete open questions.

## 2. CURRENT STATE OF RESEARCH

### *2.1 A Unified Transfer-Learning Framework*

Transfer learning is the unifying theme of this survey on low-resource ASR. The methods reviewed here differ in architecture, training protocol, and target scenario. They nonetheless share a common goal: moving useful knowledge from external sources to target tasks that lack sufficient labeled speech. Because the literature is diverse, this common logic is easy to lose. We therefore begin with a two-dimensional framework that classifies the major techniques along two complementary axes.

*Dimension I: Source of transferable knowledge*. The first axis asks where the reusable structure comes from. We distinguish three qualitatively different sources.

(a) Corpus-derived structure. Knowledge is extracted from statistical regularities in speech data. The source may be large-scale unlabeled audio, labeled speech from related languages, or unlabeled in-domain recordings. Self-supervised learning, multilingual pretraining, semi-supervised learning, and domain adaptation all belong here. The transfer vehicle is the data distribution itself.

(b) Model-encoded priors. Knowledge is already stored in the parameters of a trained model. It is reused through inheritance, constrained updating, or distillation. Layer transfer, LoRA, prompt tuning, adapters, and cross-model distillation are examples. The critical resource is the pretrained weight space.

(c) Task-structural priors. Knowledge resides in task organization and learning dynamics rather than in a single dataset or model. Meta-learning optimizes initialization and update rules across a distribution of tasks. Multi-task learning uses auxiliary objectives to shape representations that are more transferable than those learned from a single task alone. The transfer vehicle is the task distribution.

*Dimension II: Adaptation mechanism*. The second axis asks how transferred knowledge is injected into and regulated for the target task. Three canonical mechanisms appear repeatedly.

(a) Representation shaping. The feature space is shaped so that it already encodes transferable structure. Masked prediction, contrastive learning, shared encoders in multi-task learning, and cross-lingual bottleneck features all belong here.

(b) Parameter regulation. Pretrained weights are inherited, and their update is explicitly constrained. Freezing, low-rank adaptation, adapters, and regularization penalties are typical examples.

(c) Supervision augmentation. Additional training signals are synthesized from available resources. Pseudo-labeling, distillation targets, auxiliary-task losses, and TTS-generated speech all instantiate this mechanism.

*The resulting taxonomy*. The two dimensions together yield the taxonomy shown in Figure 2. The tree first branches by knowledge source. The shading of each leaf node then encodes the primary adaptation mechanism, and the node labels indicate the associated techniques. Most individual methods sit on

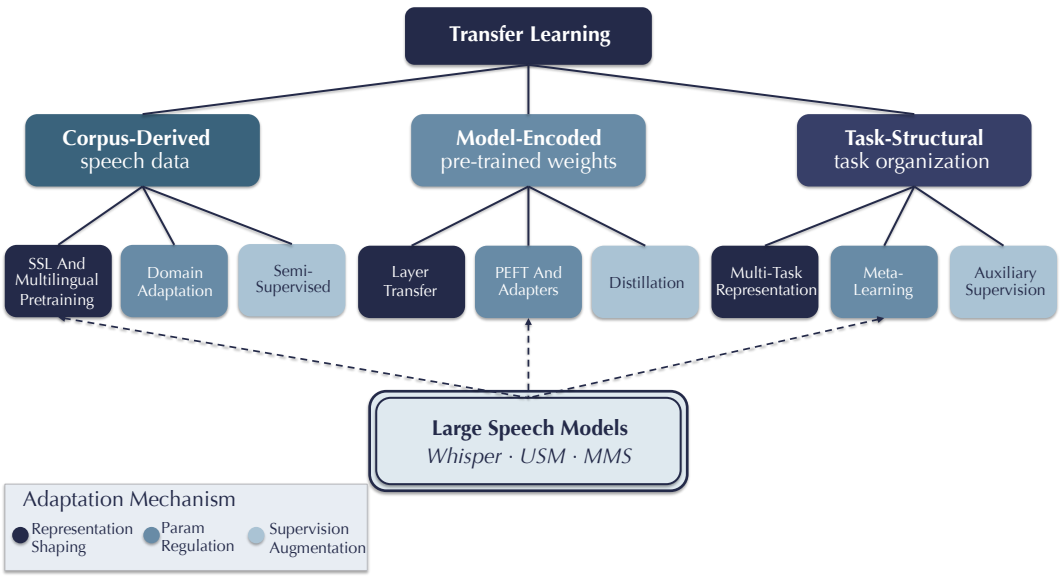



**Figure 2.** A hierarchical taxonomy of transfer-learning techniques for low-resource ASR. The tree branches by knowledge source (Dimension I); node shading encodes the primary adaptation mechanism (Dimension II). Dashed arcs indicate that large speech models draw on all three knowledge sources simultaneously.

one primary branch. This specialization gives them distinctive strengths, but it also explains why each method eventually reaches a ceiling. The strongest practical pipelines increasingly move across branches. A common example is XLSR-style pretraining followed by adapter-based fine-tuning and then semi-supervised self-training. Large multi-task speech models such as Whisper and USM go even further. They absorb multilingual corpora, encode knowledge in large parameter spaces, and support task-level adaptation. In effect, they draw on all three knowledge sources at once.

Three observations follow from this layout. First, the literature is unevenly populated. Corpus-derived and modelencoded methods dominate current work. Task-structural methods remain comparatively underexplored despite strong theoretical appeal. Second, most individual techniques excel at a single bottleneck and saturate once another bottleneck becomes active. Third, the strongest empirical systems usually chain several transfer modes end-to-end. This suggests that the techniques are better understood as composable building blocks than as mutually exclusive alternatives.

*Relationship to the theoretical analysis.* This framework complements the unified risk analysis developed later in Section 3. The present framework asks what reusable resource each method draws on and through what mechanism. The later analysis asks which measurable factor most strongly drives its gain and relates the answer to effective sample size $m_{\text{eff}}$, structural compatibility $\mathcal{C}_{\text{eff}}$, and domain shift $\Gamma$. The two views align naturally. Corpus-derived methods mainly enlarge $m_{\text{eff}}$ and reduce $\Gamma$. Model-encoded methods mainly reduce $\mathcal{C}_{\text{eff}}$. Task-structural methods often compress $\mathcal{C}_{\text{eff}}$ while also increasing $m_{\text{eff}}$. This correspondence is particularly helpful for language-specific structure. Tonal constraints and morphology-aware decoders are not separate from transfer learning. They are targeted ways of reducing the task-structure energy that generic scaling alone often fails to eliminate.

*Guide to the subsequent sections.* The taxonomy organizes the rest of the paper. Sections 2.2 and 2.3 examine corpusderived and model-encoded transfer through pretraining and fine-tuning. Section 3 provides a cross-cutting theoretical and empirical analysis. Section 4 turns to task-based transfer, large-scale multi-task speech models, and corpus expansion strategies. Throughout the paper, the reader can locate a method within the taxonomy and immediately see both its primary transfer mode and its relation to the rest of the field.

### 2.2 Pre-training

#### 2.2.1 Cross-Lingual Self-Supervised Representation Learning

*Universal representation extraction and adaptation.* Cross-lingual self-supervised learning (CL-SSL) is now the dominant pretraining paradigm for low-resource ASR. Its basic idea is straightforward. A model first learns acoustic structure from large multilingual unlabeled corpora, and only then is it adapted to the target language with limited supervision. This strategy extends the older multilingual tradition of shared bottleneck features and shared hidden layers [18-19]. It also fits naturally with end-to-end CTC, attention, and Transformer architectures, where language IDs, gating, or adaptive decoding can reduce cross-lingual interference in few-shot and zero-shot settings [20-21]. The main appeal of CL-SSL is that it offers a

better balance between sample efficiency, robustness, and cross-lingual reuse than purely supervised multilingual training.

At the modeling level, CL-SSL learns language-agnostic but adaptable acoustic structure through masked prediction or contrastive learning. The wav2vec family is the canonical example, together with its cross-lingual extensions [22-23]. The XLSR architecture, one representative cross-lingual extension of this SSL paradigm, is illustrated in Figure 3. The major representative models are summarized in Table 1. In wav2vec 2.0, the encoder and context network are coupled through masked modeling and a contrastive objective [24]:

$$\mathcal{L}_{\text{wav2vec2}} = -\sum_{t} \log \frac{\exp\left(\sin(g(\mathbf{z}_t),\ \mathbf{q}_t)\right)}{\sum_{n \in \mathcal{N}_t} \exp\left(\sin(g(\mathbf{z}_t),\ \mathbf{q}_n)\right)} + \lambda \parallel \mathbf{z} \parallel_2^2$$

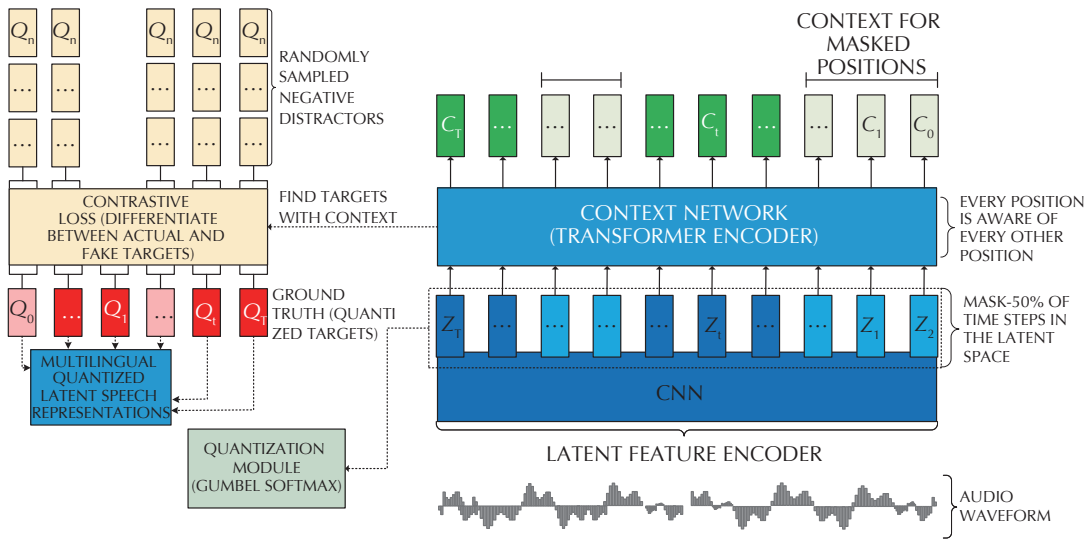



**Figure 3.** Illustration of the XLSR Model Architecture.

Later work addressed known weaknesses of this objective, especially false negatives and sensitivity to local perturbations, through reweighting, pruning, or architectural coupling. HuBERT follows a different route. It replaces explicit negative sampling with masked prediction of discrete labels obtained by unsupervised clustering, and in very low-resource settings it often provides more stable hierarchical units [25].

When self-supervised pretraining is scaled to multilingual corpora, the XLSR family and its variants add a second ingredient: language balancing. Without such balancing, high-resource languages dominate the objective. XLSR-style sampling counters this effect through inverse-size or temperature-based weighting [35, 41-42]:

$$\ell \sim \text{Multinomial}(\alpha_1,\ \ldots,\ \alpha_L), \quad \mathbf{x}^{\ell} \sim \mathcal{P}_{\ell}, \quad \alpha_{\ell} = \frac{\gamma}{(|\mathcal{D}_{\ell}| + \epsilon)^{\beta}}$$

**Table 1.** Summary of Self-Supervised Models.

| Experimental Model | Dataset | Baseline | | Experimental | |
|---|---|---|---|---|---|
| | | Performance | Parameters | Performance | Parameters |
| APC [26] | LibriSpeech | WER: 61.91% | 5.5M WER: | 64.66% | 5.5M |
| LF-MMI [27] | MGB | WER: 29.4% | 16.7M | WER: 24.0% | 16.7M |
| FNIE [28] | CommonVoice | WER: 5.76% | ~95M | WER: 4.936% | ~95M |
| EDCL [29] | CommonVoice | WER: 13.68% | 94M | WER: 12.36% | 94M |
| Wav2vec-S [30] | SWBD | WER: 19.9% | 95M | WER: 14.2% | 95M |
| vq-Wav2vec [31] | MuST-C En-Fr | BLEU: 27.8 | 95M | BLEU: 28.6 | 8M |
| Data2vec 2.0 [32] | Librispeech-Large | WER: 4.6 | 318M | WER: 4.3 | 318M |
| HUBERT [33] | Libri-light | WER: 15.7% | 95M | WER: 15.0% | 95M |
| IPA-informed HuBERT [34] | MLS | CER: 10.6% | 300M | CER: 8.55% | 317M |
| XLSR-53 [35] | HYKIST | WER: 25.4% | 100M | WER: 23.9% | 103M |
| XLS-R [36] | CORAAL | WER: 24.8% | 30M | WER: 12.7% | 317M |
| WavLM [37] | OpenASR21 | WER: 57.1% | 300M | WER: 46.5% | 300M |
| DRAFT [38] | OGI test | CER: 2.5% | 95M | CER: 2.05% | 108.7M |
| LABERT [39] | I-CUBE Large | WER: 11.28% | 340M | WER: 9.53% | 340M |
| EFFUSE [40] | Totonac | WER: 45.2 | 900M | WER: 46.9 | 624M |

This simple design choice has practical consequences. It makes multilingual pretraining less biased toward a few dominant languages and increases the chance that rare phonetic or prosodic patterns remain visible during adaptation.

*Tone- and morphology-aware transfer*. Generic CL-SSL is powerful, but it does not automatically preserve every linguistically decisive contrast. This limitation is most visible in tonal languages and in languages with rich or agglutinative morphology. In tonal languages, invariance learning is beneficial only up to a point. It should suppress speaker-specific pitch range and channel variation, but it must not collapse lexical tone itself. A practical response is to expose F0 or pitch-contour information to the encoder through auxiliary embeddings, F0-conditioned layers, or explicit tone heads [43-48]. A generic formulation is

$$h_t = \mathrm{End}([x_t;\ e_t^{\mathrm{F0}};\ e_t^{\mathrm{tone}}]), \quad \mathcal{L}_{\text{tone-aware}} = \mathcal{L}_{\mathrm{ASR}} + \lambda_{\mathrm{tone}}\mathcal{L}_{\text{tone-ctr}} + \mu_{\mathrm{sd}}\mathcal{R}_{\mathrm{sandhi}}$$

Here $\mathcal{L}_{\text{tone-ctr}}$ keeps tone minimal pairs separated in representation space, and $\mathcal{R}_{\text{sandhi}}$ penalizes outputs that violate known tone-sandhi mappings under their local context. Conceptually, these additions are orthogonal to CL-SSL rather than alternatives to it. CL-SSL provides the broad acoustic prior. Tone-aware constraints decide which suprasegmental distinctions the model must not erase.

A parallel issue appears in morphologically rich and agglutinative languages. Frequency-driven BPE is convenient, but it often merges across true morpheme boundaries. In low-resource conditions, this can inflate sparsity because many inflected forms appear only a few times. A common remedy is to reconstruct the output space with morphology-aware units: stem-affix factorization, linguistically guided subword segmentation, pronunciation attributes, or auxiliary heads that predict morpheme boundaries and inflectional features [49-54]. One compact form is

$$\mathcal{L}_{\text{morph-aware}} = \mathcal{L}_{\text{CTC/attn}} + \lambda_{\text{seg}}\mathcal{L}_{\text{morph-seg}} + \lambda_{\text{feat}}\mathcal{L}_{\text{morph-tag}}$$

where the auxiliary losses predict morpheme boundaries and morpho-syntactic bundles such as case, tense, number, or voice. The practical trade-off is clear. Plain BPE offers simplicity and broad compatibility. Morphology-aware units impose more design effort, but they often improve transfer because updates on one form can generalize to many related forms.

*Downstream integration.* CL-SSL does not replace multilingual learning. It reshapes it. Shared backbones with languagespecific heads, block-softmax variants, and multi-task cross-entropy losses remain useful as lightweight adaptation layers on top of self-supervised representations [19]. Architectures that emphasize temporal or local structure, such as CNN-RNN hybrids, also benefit once strong self-supervised features are available [55]. The same is true of sparsified or languagerouted subnetworks, language feature vectors, and gating modules, all of which address the central trade-off between sharing and specialization [56].

On the decoding side, end-to-end multilingual ASR and speech translation systems often combine self-supervised front-ends with CTC/attention objectives, external language models, or WFST-style decoding [57]. This combination is especially useful in code-switching and in morphologically complex languages. The front-end extracts reusable acoustic units. The output side then decides how much language-specific structure is needed. This division of labor is one reason why CL-SSL remains effective even when the target language differs strongly from the pretraining pool.

*Optimization and robustness.* Engineering practice has also converged on a small set of robust optimization strategies. TERA-style perturbation-based objectives improve robustness to noise and recording variation [58]. Cyclical learning rates and related optimizer tricks stabilize convergence [59]. Feature- and label-side augmentation further enlarge the effective sample size under low-resource conditions [60]. These choices may look secondary, but in low-resource settings they often determine whether the theoretical transfer potential of CL-SSL is realized in practice.

*Self-training.* A second stage is often needed after CL-SSL: semi-supervised self-training on unlabeled target-domain data. This step extends the representational advantage of CL-SSL into domain-specific

supervision. The core mechanism is simple. The model generates pseudo-labels for unlabeled target speech, filters them by confidence, and then retrains on the accepted set. In practice, this is one of the most dependable "second-stage" strategies for low-resource ASR [30]. Backend rescoring and model ensembling can further improve robustness on the decoding side [61]. Parameter fusion, pruning, and sparsification reduce inference cost and make deployment more realistic [62]. Deep multilingual backbones can even reduce overfitting rather than increase it when the target data are small, especially when residual or VGG-style structures are used in joint training [63].

The influence of CL-SSL also extends beyond core ASR. Shared self-supervised representations have been reused in multilingual keyword spotting, emotion recognition, sign language processing, and code-switching [64]. In child speech, elderly speech, and dialectal speech, CL-SSL is often paired with active learning to prioritize high-value annotations. Current best practice therefore looks less like a single model and more like a recipe: multilingual self-supervised pretraining, lightweight target adaptation, selective output-side specialization, and confidence-filtered self-training [29, 42, 49, 65-66].

#### *2.2.2 Unsupervised/Semi-Supervised Acoustic Unit Discovery*

This line of work converts unlabeled speech into usable supervision. The pipeline usually has three stages. The first stage learns transferable representations. The second stage stabilizes pseudo-label quality on the target language or domain. The third stage reduces uncertainty in the output space through subwords, lexicons, or decoding constraints. These stages are successive rather than interchangeable, and they are often most effective when combined with CL-SSL rather than used in isolation.

The unit-discovery route inherits ideas from earlier multilingual supervised ASR. Within DNN-HMM, CTC, or attention-based systems, researchers have used phoneme confusion matrices, JSD divergence, and hierarchical clustering to build universal phoneme inventories or cross-lingual subword spaces [50]. In morphologically rich languages, lexicon induction and statistical subword construction help reduce OOV rates and sparsity. Under even lower supervision, a multilingual ASR teacher can first generate rough pseudo-transcriptions, after which K-means or DPGMM clustering provides more stable weak labels for the target language [67]. At the extreme unsupervised end, wav2vec-U maps audio directly to phoneme sequences through adversarial matching and HMM self-training [68]. This line of work remains challenging, but it reveals how far one can go when transcriptions are unavailable but text priors still exist.

The semi-supervised stage is more mature and more broadly useful. Its main benefit is statistical: a small amount of trusted labeled data calibrates the decision boundary, while large volumes of unlabeled speech reduce variance. This logic only works, however, if pseudo-label quality is controlled. A representative objective is the one used in slimIPL:

$$\mathcal{L} = \mathcal{L}_L + \lambda \mathcal{L}_U = -\sum_{(x,y)\in L} \log p_\theta(y \mid x) - \lambda \sum_{x \in U} \log p_\theta(\hat{y} \mid x)$$

where $\hat{y}$ is obtained by decoding the unlabeled speech with the current model [69]. The empirical lesson is consistent across languages. Northern Sotho gains from moderate noise augmentation plus self-training [70]. Swahili benefits from low-shot multilingual transfer with shared hidden layers [71]. More recent pipelines based on self-supervised large models and iterative pseudo-labeling on target-domain speech have proven effective on multilingual corpora such as VoxPopuli [72].

Two practical points are worth stressing. First, these methods are not substitutes for representation learning. In current best practice, unsupervised or self-supervised pretraining provides the base representation, semi-supervised self-training expands target-side supervision, and subword or unit discovery stabilizes the output space when labels are extremely scarce. Second, the value of unlabeled data is not monotonic. Once pseudo-label noise accumulates faster than usable supervision, gains flatten and can reverse. This is why modern pipelines add confidence filtering, uncertainty-based rejection, or teacher rescoring [73-74].

*From representation learning to supervision induction.* CL-SSL and unsupervised or semi-supervised unit discovery solve different parts of the low-resource problem. CL-SSL mainly reduces $\mathcal{C}_{\text{eff}}$ by building a stronger acoustic prior, and it can also reduce $\Gamma$ when multilingual or in-domain pretraining is broad enough. Unit discovery and self-training mainly increase $m_{\text{eff}}$ by converting unlabeled target speech into supervision. Their ceiling is therefore set less by backbone capacity than by pseudo-label quality. This distinction becomes sharper in tonal and morphologically rich languages. If tone categories or morpheme boundaries are already preserved in the representation and output space, self-training can amplify them. If those distinctions were blurred at the representation stage, pseudo-labeling may simply reinforce the wrong decision boundary. In practice, the most reliable route is usually a CL-SSL backbone, followed by language-aware output design and confidence-filtered self-training [30, 49, 50, 68, 72].

### 2.3 Fine-tuning

#### 2.3.1 PEFT

*Technical overview.* Parameter-efficient fine-tuning asks a focused question: how much of a pretrained model must be updated for a new low-resource language? This question matters because full fine-tuning is flexible, but it is also expensive and unstable when target labels are scarce. A convenient unified objective is

$$\min_{\theta}\left(\mathcal{L}_{\text{ASR}}(\mathcal{D}_t;\ \theta)+\lambda\sum_i \Omega(\theta_i,\ \theta_i^s)\right)$$

where $\Omega$ penalizes deviations from source parameters, as in EWC- or MAS-style regularization [75]. This formulation covers a spectrum of methods, including full fine-tuning, restricted layer transfer, and structured low-rank or bottleneck adaptation.

The empirical picture is now fairly consistent. Self-supervised or multilingual pretraining supplies cross-lingually transferable acoustic abstractions, and even a small amount of target supervision can then deliver a substantial reduction in WER or CER [76]. When the target data are extremely scarce, pre-finetuning

or multi-stage transfer often smooths the final adaptation step [77]. Data-side controls such as language-similarity weighting and curriculum learning further improve sample efficiency by changing which examples dominate the gradient early in training [78]. On the decoding side, joint CTC/attention training and dual-branch inference remain robust choices, especially when combined with an external language model [79]. These patterns have been reported across a wide range of languages and settings, including Indian languages, Indonesian, Tibetan, Russian, African languages, Kannada dialectal ASR, Bengali multimodal systems, and agglutinative Central Asian languages [80-83].

*Deployment*. In practical systems, the most common strategy is to freeze what appears broadly reusable and to update only what appears task-specific. This usually means keeping the lower acoustic front-end fixed, then adapting a small number of middle or upper layers [46]. When paired with strong self-supervised pretraining, this restricted update pattern often remains effective even under severe target scarcity [84]. Layer aggregation, layer dropping, early exit, and sequence downsampling further reduce inference cost. Learnable interpolation across layers can also mitigate overconfidence in higher layers [85]. Similar observations recur in work on Mongolian, Malay, Urdu, Persian, and Afrikaans [45, 86-89].

Adapters make this logic more explicit. By inserting bottleneck modules and freezing the backbone, one can often update only 1%–10% of the parameters while approaching or sometimes surpassing full fine-tuning. This is especially attractive in long-tail, multi-dialect, or multi-domain settings [90]. Adapters can also be combined with language-similarity weighting, meta-learning, or binary routing schemes, which helps the model select which source knowledge to reuse for which target condition. When combined with self-supervised representations, middle-to-upper-layer adapters are often sufficient for cross-lingual and cross-domain transfer [91]. Coupling them with multi-task CTC-attention heads further balances robustness and plasticity [92].

*Language-specific adaptation under a small budget*. PEFT becomes more informative when the trainable budget is allocated with the target language in mind. In tonal languages, the most useful updates are often not distributed uniformly across the network. Mid-to-upper encoder layers and decoder-side projections are the places where pitch, segmental evidence, and text constraints are fused. Tone-aware LoRA or adapter modules placed in these regions can therefore be more effective than uniform low-rank adaptation [43, 45-46]. In morphologically rich languages, the opposite lesson often holds: the output side matters disproportionately. Small modules attached to the tokenizer, predictor, or decoder projection can recover more than broad acoustic adaptation if the main problem is stem-affix composition rather than acoustic mismatch [49, 50, 91-92]. In this sense, PEFT does not replace language-aware design. It provides the budget discipline within which language-aware design becomes practical.

*Operational trade-offs*. The PEFT family contains three distinct operating points. Full fine-tuning gives the highest adaptation freedom, but it also carries the greatest risk of overfitting and forgetting. Layer transfer decides *where* updates should occur and is often the cleanest way to stabilize training. Adapters and LoRA decide *how much* of the shared space should be modified and therefore give the best parameter efficiency. In terms of Eq. (13), all three methods mainly act on $\mathcal{C}_{\text{eff}}$ by controlling how far the posterior is

allowed to move away from the inherited prior. Their trade-off is equally clear. The smaller the adaptation subspace, the better the stability and deployment cost, but the lower the ceiling under large domain or label-space shift.

### *2.3.2 Domain Adaptation*

Domain adaptation becomes central when the main obstacle is mismatch rather than the absolute lack of speech. The mismatch may be acoustic, demographic, linguistic, or textual. Accent, dialect, child speech, articulatory disorder, device shift, scenario shift, and genre shift all fall into this category [93-96]. A broad body of work shows that these mismatches matter across low-resource languages such as European Portuguese, Vietnamese, Amharic, Uyghur, Tujia, and Tashlhiyt [97-102].

Two technical routes dominate. The first works in parameter space. Adapters, constrained updates, or residual injections modify the model only where domain-specific corrections are needed [96]. The second works in representation space. Multi-stream architectures, domain discriminators, and cross-modal fusion attempt to learn features that are domaininvariant but still task-discriminative [103-104]. This is why adversarial alignment and multi-task adaptation have become common. They directly target the divergence between training and test distributions.

Large self-supervised backbones have strengthened this line of work. Models such as wav2vec 2.0, HuBERT, and XLS-R provide a transferable low-level representation, after which relatively light adaptation can already yield robust gains [105]. When the shift affects the ontology of the output space, cross-lingual subspace constraints based on the IPA or shared phoneme subsets can further reduce the reconstruction cost of the decoder [106]. Uncertainty-based pseudo-label filtering is another recurrent pattern. It limits the damage caused by noisy self-training when unlabeled target speech is available [107]. If source and target diverge too strongly, a safer choice is to keep the pretrained encoder as a fixed feature extractor and rebuild only the decoder or language model in the target domain. External LM or WFST fusion can then recover textual priors that the acoustic adaptation alone cannot supply [108].

A point that is easy to overlook is that domain adaptation must preserve the right information, not just suppress domain markers. This is especially important in tonal languages. Pitch normalization and adversarial invariance can help remove speaker-specific variation, but if they are applied without care they can also erase lexical tone. A better strategy is to separate speaker-pitch normalization from tone-preserving F0 cues, and, when possible, to align domains with tone-aware critics or prosody heads rather than tone-blind adversarial objectives [43-45]. A similar issue appears in morphology-rich languages. Acoustic adaptation alone is often insufficient if the true shift sits in the output space, for example in productive inflection or word formation. In such cases, domain adaptation should be paired with morphology-aware segmentation or decoding rather than being treated as a front-end fix only.

*What this route buys*. Relative to PEFT, domain adaptation spends more of the optimization budget on reducing $\Gamma(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$ than on enlarging the adaptation subspace itself. This is why it often yields fast early gains under accent, device, and scenario mismatch. It is also why it plateaus earlier when the

remaining mismatch is in the label space rather than the acoustic space. In practice, domain adaptation works best when unlabeled target speech is available, when the source prior is already strong, and when output-side corrections are added for phenomena such as lexical tone or productive morphology.

### *2.3.3 Cross-Model Distillation*

*Scope of applicability.* Distillation is most useful when the target model is small, the target labels are sparse, or deployment cost is strict. Instead of asking the student to infer the whole decision structure from hard labels alone, distillation transfers that structure from one or more teachers. At the output level, temperature-scaled KL and hard-label cross-entropy are the standard combination. Hard labels anchor the decision boundary, whereas soft labels communicate relative class structure [109]. Under low-resource conditions, this can improve calibration and sample efficiency, especially when combined with linearization constraints such as Mixup.

Representation-level distillation attacks a different problem. It aligns intermediate states between networks and thereby shapes a more language-agnostic acoustic subspace. Typical cases include hidden-state alignment between TDNN-LSTM and TDNN systems and speech-model compression without access to very large multilingual corpora [110]. Sequencelength mismatch and CTC blanks make this difficult, but shrinkage, nearest-neighbor interpolation, and learnable projections can substantially reduce the alignment problem. DistilXLSR and LightHuBERT are representative examples. The former emphasizes cross-lingual robustness. The latter emphasizes deployment flexibility and energy efficiency [111].

*Clarification of technical principles.* Multi-teacher distillation is not just averaging several teachers. It is a mechanism for deciding who should teach what, and where. In multilingual training, language-specific teachers can provide posteriors or hidden representations to a shared student, while weighting and regularization reduce conflicts among teachers [112]. Structural perturbations of language-dependent modules can further enlarge teacher diversity and improve the final student [113]. Cross-modal distillation extends the same idea. By exploiting the duality between TTS and ASR, one can generate pseudo data and then use distillation to control reliability, which is particularly useful under minutes-to-hours annotation budgets [114]. Similar ideas have also been applied to spoken language understanding and deployable end-to-end systems [115].

A practical warning follows from this literature. Continued pretraining on the target domain often introduces a stability-plasticity trade-off. Self-distillation mitigates this by constraining both representation and parameter drift between the initial model and the continued-pretraining teacher [116]. The same lesson applies to language-specific structure. If a teacher already collapses rare tone categories or morphology, the student inherits that blindness. Distillation is therefore strongest when the teacher's structure is genuinely better than the student's, not merely larger.

*Reading the fine-tuning family as a whole.* PEFT, domain adaptation, and distillation should be viewed as complementary controls over Eq. (13). PEFT mainly regulates posterior movement and therefore $\mathcal{C}_{\text{eff}}$. Domain adaptation acts most directly on $\Gamma$. Distillation changes the effective supervision seen by the

student and therefore reshapes the empirical-risk term. None dominates in all regimes. PEFT is efficient, but it can saturate under severe shift. Domain adaptation is effective when mismatch dominates, but it may leave decoder-side sparsity unresolved. Distillation can outperform both when the teacher is substantially better, yet it can also fail when the teacher transfers the wrong structure. This is why hybrid pipelines—for example, SSL pretraining plus LoRA/adapters, together with tone- or morphology-aware heads and target-domain self-training or distillation—are increasingly more realistic than any single method in isolation [75, 91-92, 103, 109, 116].

## 3. CURRENT TECHNICAL BOTTLENECKS AND CHALLENGES

### *3.1 Overview of the technology*

The preceding sections show that contemporary low-resource ASR has largely converged on a pretrain-adapt workflow, yet that apparent convergence conceals a more consequential difference: the eight mainstream techniques do not solve the same bottleneck. Multilingual pre-training and self-supervised learning primarily build transferable structure before the target task is even seen; domain adaptation and un-/semi-supervised learning try to unlock target-side information that remains latent when labels are scarce; cross-model distillation, model-layer transfer, PEFT, and adapters regulate how previously acquired knowledge is injected into a small-data regime. In practice, then, their gains arise from different levers. What varies across papers is therefore not only the absolute performance level, but the binding constraint itself: some settings are limited by insufficient upstream diversity, some by the amount of target annotations, some by domain mismatch, and others by the narrow adaptation budget that the model can safely tolerate.

This distinction matters for this chapter. Even when the same nominal paradigm is used, the target language, corpus quality, task definition, annotation density, model scale, and decoding conditions usually differ substantially across studies. Individual results are informative, but their juxtaposition rarely yields an immediately comparable picture. The aim of this chapter is therefore to recover a set of operational regularities.

### *3.2 A normalization method for heterogeneous evidence*

To make such comparison possible, we introduce **Difficulty-Adjusted Gain (DAG)**, an effective cross-task heterogeneous normalization method designed for cross-paper synthesis rather than as a replacement for task-specific metrics. The motivation is straightforward: raw WER, PER, BLEU, Accuracy, UAR, and F1 are not directly comparable because they differ in scale, direction, and saturation behavior. A reduction from 10% to 8% WER is not equivalent, in difficulty, to an increase from 0.88 to 0.90 accuracy, even if the absolute numerical change looks similar.

DAG proceeds in three steps. First, every reported result is rewritten as an error rate $e$: lower-is-better metrics such as WER and PER are used directly, while higher-is-better metrics such as Accuracy, BLEU, UAR, and F1 are mapped to 1 – score. Second, we apply a logit transform to place all results on a common real-valued scale.

$$\text{logit}(e) = \ln \frac{e}{1-e}$$

Third, we measure the gap to a task-specific scratch anchor baseline, with a lightweight language-difficulty correction when the target language is not English. The resulting score is

$$\text{DAG} = \text{logit}(e^{\text{adj}}_{\text{anchor}}) - \text{logit}(e_{\text{system}})$$

A positive DAG therefore means that the reported system outperforms the anchor, and a larger value indicates a gain that remains meaningful even after differences in task scale and language difficulty have been normalized.

To keep this reference system transparent, we built 17 lightweight from-scratch anchors covering the full task set. Only the core information is retained here, since the role of the anchor is comparative rather than procedural. Table 2 reports the compact anchor configuration across the 17 tasks

**Table 2.** Compact anchor configuration across the 17 tasks.

| Task | Group | Fam. | Dataset | Train hrs | Metric | Anchor |
|---|---|---|---|---|---|---|
| ASR | $\alpha$ | A | LibriSpeech | 10.0 | WER ↓ | 0.52 |
| LVCSR | $\alpha$ | A | same as ASR | 10.0 | WER ↓ | 0.55 |
| WR | $\alpha$ | A | same as ASR | 10.0 | WER ↓ | 0.51 |
| PR | – | A | TIMIT | 3.2 | PER ↓ | 0.19 |
| ST | $\gamma$ | B | CoVoST2 En → De | 10.0 | BLEU ↑ | 0.03 |
| AST | $\gamma$ | B | same as ST | 10.0 | BLEU ↑ | 0.04 |
| KWS | $\beta$ | D | Speech Commands v2 | 1.67 | Accuracy ↑ | 0.89 |
| KS | $\beta$ | D | same as KWS | 1.67 | Accuracy ↑ | 0.90 |
| SCR | $\beta$ | D | same as KWS | 1.67 | Accuracy ↑ | 0.54 |
| SLU | – | C | SLURP | 10.0 | Accuracy ↑ | 0.88 |
| IC | – | C | Fluent Speech Commands | 9.8 | Accuracy ↑ | 0.84 |
| SER | – | C | IEMOCAP | 3.0 | UAR ↑ | 0.53 |
| SID | – | E | VoxCeleb1 | 10.0 | Accuracy ↑ | 0.76 |
| DID | – | E | Common Voice EN | 2.5 | Accuracy ↑ | 0.63 |
| LID | – | E | VoxLingua107 | 1.67 | Accuracy ↑ | 0.72 |
| ER | – | C | SLURP (entity) | 10.0 | F1 ↑ | 0.42 |
| MDD | – | A | L2-ARCTIC | 20.3 | F1 ↑ | 0.37 |

All anchors use three seeds [13, 17, 23], speaker-disjoint splits where available, dev-only model selection, and the median test score as the final anchor. Shared runs: $\alpha$ = ASR/LVCSR/WR, $\beta$ = KWS/KS/SCR, $\gamma$ = ST/AST.

These anchors are implemented with five lightweight model families so that the reference points remain stable while staying close to realistic speech pipelines. Family A is a compact Conformer-CTC model for ASR-like tasks; Family B uses a Conformer encoder with a small Transformer decoder for ST/AST; Family C attaches task-specific classification or tagging heads to the same Conformer encoder for SLU, IC, SER, and ER; Family D is a compact ResNet-15 for KWS/KS/SCR; and Family E is an ECAPA-TDNN for speaker-, dialect-, and language-identification tasks. Across families, we retain only the core commonalities that matter for comparison: 16 kHz audio, 80-dim log-Mel features, SpecAugment, validation-based early stopping, and three-seed median reporting.

On top of DAG, we extract 11 factors to analyze what actually drives cross-lingual transfer performance across the literature. Table 3 lists the eleven factors used in the DAG regression analyses

**Table 3.** Eleven factors used in the DAG regression analyses.

| ID | Factor | Operational meaning | Transform |
|---|---|---|---|
| $f_1$ | Upstream pretraining hours | Total hours of speech used before target-task adaptation, including multilingual supervised training or SSL pretraining | $\log_{10}$ |
| $f_2$ | Backbone parameters | Total backbone parameter count (in millions) | $\log_{10}$ |
| $f_3$ | Trainable parameters | Parameters actually updated during target-task adaptation (in millions) | $\log_{10}$ |
| $f_4$ | Trainable ratio | Relative adaptation magnitude, i.e., $f_3/f_2$ | logit |
| $f_5$ | Target labeled hours | Labeled speech hours in the target language | $\log_{10}$ |
| $f_6$ | Target unlabeled hours | Unlabeled speech hours in the target language | $\log_{10}$ |
| $f_7$ | Number of pretraining languages | Count of languages seen upstream | $\log_{10}$ |
| $f_8$ | Language offset | Difficulty offset derived from cross-lingual ASR performance differences | raw |
| $f_9$ | Labeled/unlabeled ratio | Relative density of supervision, i.e., $f_5/f_6$ | $\log_{10}$ |
| $f_{10}$ | PEFT rank/bottleneck size | LoRA rank or adapter bottleneck dimension | $\log_2$ |
| $f_{11}$ | Teacher error rate | Error rate of the distillation teacher | logit |

Most extensive variables are entered in log space because the literature spans several orders of magnitude. Factors $f_{10}$ and $f_{11}$ are structurally applicable only to PEFT/adapter and distillation settings, respectively.

The point of DAG is therefore not to force all studies into an artificial single benchmark, but to make their evidence commensurable enough for patterns to emerge.

### 3.3 Cross-study synthesis and discussion

With the anchors and factors in place, the broader literature can be organized in a way that is comparable without flattening away the differences that actually matter. The large synthesis in Table 4 should thus be read less as a leaderboard than as a map of operating regimes, dominant drivers, and typical data budgets across the field.

The broad patterns in Table 4 become more precise when the underlying factor structure is analyzed directly, as shown in Figure 4.

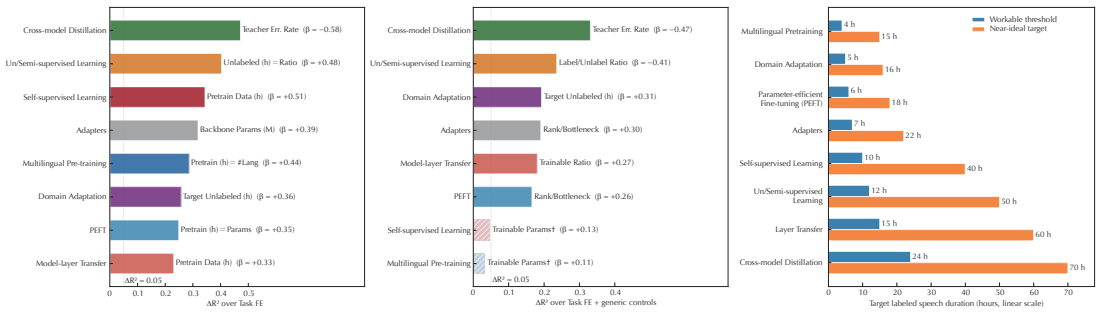
Panels (a) and (b) are especially useful because they show that the field is not governed by a single universal scaling law. In the dominant-factor analysis, multilingual pre-training is mainly organized by upstream pretraining hours $f_1$ ($\Delta R^2 \approx 0.287$, $\beta \approx +0.438$), but the number of pretraining languages $f_7$ is almost equally informative. This is why multilingual studies are better understood through the joint lens of scale and language coverage rather than scale alone. Selfsupervised learning is even more clearly scale-sensitive, with $f_1$ providing the strongest explanatory power ($\Delta R^2 \approx 0.344$, $\beta \approx +0.512$). By contrast, cross-model distillation is driven primarily by teacher error rate $f_{11}$ ($\Delta R^2 \approx 0.472$, $\beta \approx -0.579$); domain adaptation by target unlabeled hours $f_6$ ($\Delta R^2 \approx 0.258$, $\beta \approx +0.362$); un-/semi-supervised learning jointly by target unlabeled hours $f_6$ and labeled/unlabeled ratio $f_9$ ($\Delta R^2 \approx 0.404$ and 0.389, respectively); model-layer transfer by upstream hours $f_1$; PEFT by upstream hours $f_1$ with backbone size $f_2$ close behind; and adapters by backbone size $f_2$ ($\Delta R^2 \approx 0.319$, $\beta \approx +0.391$).

Panel (b) sharpens the picture by asking which factor remains explanatory after generic scale effects are discounted. Multilingual pre-training and self-supervised learning show no equally strong technique-exclusive driver; their gains appear to arise from broad upstream quality itself, rather than from a narrow secondary knob. The remaining six techniques do show clearer specific drivers: distillation on teacher quality $f_{11}$ ($\Delta R^2 \approx 0.331$); domain adaptation on target unlabeled hours $f_6$ ($\Delta R^2 \approx 0.193$); un-/semi-supervised learning on the labeled/unlabeled ratio $f_9$ ($\Delta R^2 \approx 0.236$, $\beta \approx -0.407$); model-layer transfer on trainable ratio $f_4$ ($\Delta R^2 \approx 0.181$); and both PEFT and adapters on the adaptation bottleneck $f_{10}$ ($\Delta R^2 \approx 0.166$ and 0.191, respectively). This is why the sorting key in Table 4 is technique-specific rather than uniform. When the literature reveals a directly observable operational axis, that axis is used in the table. When the deeper mechanism is less visible, panel (b) clarifies what most directly explains the gain.

Panel (c) translates these statistical regularities into practical target-label regimes. It is best read as an operating map rather than as a rigid threshold chart. According to the operating points summarized in the panel, multilingual pre-training becomes workable at around 4 h of target labels and approaches a near-ideal regime at around 15 h. Domain adaptation becomes workable at around 5 h and approaches its near-ideal regime at around 16 h. LoRA-style PEFT becomes workable at around 6 h and approaches its near-ideal regime at around 18 h, while adapters become workable at around 7 h and approach their near-ideal regime at around 22 h. Self-supervised learning requires a larger target-label budget, becoming workable at around 10 h and approaching a near-ideal regime at around 40 h. Un-/semi-supervised learning becomes workable at around 12 h and approaches its near-ideal regime at around 50 h. Model-

**Table 4.** A Comprehensive Overview of Low-Resource Automatic Speech Recognition Technologies.

| Challenge | Technique | Task | Avg. tgt hrs | Related work | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Empirical Performance Trends (DAG Score) | | | | | | | | Sorting criterion |
| Audio scarcity | Multilingual pre-training | ASR, SER | 8 h | 0.226[117] 1.026[123] | 0.352[118] 1.150[124] | 0.367[119] 1.255[21] | 0.521[120] 1.400[55] | 0.677[121] 1.483[125] | 0.790[122] | 0.835[61] | 0.887[65] | 0.983[20] | f1 |
| | Self-supervised learning | ASR, PR | 97 h | 0.262[26] 1.224[37] | 0.426[28] 1.344[32] | 0.533[29] | 0.573[126] | 0.815[25] | 1.036[127] | 1.069[128] | 1.108[33] | 1.215[34] | f1 |
| | Un-/semi-supervised learning | ASR, PR | 78.2 h | 0.643[129] | 0.676[130] | 0.734[131] | 1.058[132] | 1.140[133] | 1.254[74] | 1.541[73] | 1.628[134] | | f6 |
| Text scarcity | Cross-model distillation | ASR | 65.5 h | –0.082[135] | 0.194[113] | 0.365[136] | 0.844[111] | 0.913[116] | | | | | f11 |
| | Domain adaptation | ASR, ST, SCR | 3.6 h | 0.214[137] 0.631[141] | 0.244[93] 0.636[142] | 0.365[100] 0.721[143] | 0.436[138] 0.756[144] | 0.479[139] 0.824[103] | 0.506[140] 0.824[145] | 0.514[95] 0.879[146] | 0.522[106] 1.042[147] | | f6 |
| | Model-layer transfer | ASR, AST, KWS | 36 h | 0.080[86] 0.954[154] 1.982[84] | 0.177[148] 1.165[88] | 0.188[89] 1.412[155] | 0.383[149] 1.512[156] | 0.707[150] 1.611[157] | 0.726[151] 1.630[158] | 0.805[152] 1.638[159] | 0.812[153] 1.884[160] | | f1 |
| Accent variation | PEFT | ASR, SLU, IC | 4 h | 0.098[161] 0.864[78] | 0.207[162] 0.985[168] | 0.219[163] 1.081[169] | 0.279[80] 1.186[170] | 0.590[164] 1.372[171] | 0.705[165] 1.512[172] | 0.807[166] 1.536[81] | 0.846[167] 1.628[75] | | f1 |
| | Adapters | ASR, WR | 10 h | 0.634[173] | 0.670[174] | 0.781[175] | 0.788[176] | 0.925[91] | 0.967[177] | 1.007[92] | 1.108[178] | | f2 |
| | Domain adaptation | ASR, DID, LVCSR | 5.3 h | 0.211[99] 0.616[104] | 0.237[98] 0.710[181] | 0.292[105] 0.806[182] | 0.301[102] 0.807[96] | 0.461[101] 0.900[107] | 0.471[179] | 0.489[97] | 0.554[108] | 0.596[180] | f6 |
| Morphological richness | Multilingual pre-training | ST, KWS | 2 h | 0.428[51] 1.365[60] | 0.558[183] 1.480[52] | 0.593[63] 1.524[186] | 0.642[19] 1.562[53] | 0.741[18] | 0.816[184] | 0.824[185] | 1.146[54] | 1.213[57] | f1 |
| | Self-supervised learning | AST, SLU | 58.8 h | 0.502[187] 1.251[40] | 0.549[39] 1.444[38] | 0.609[188] | 0.667[27] | 0.751[35] | 0.834[189] | 0.973[36] | 1.079[190] | 1.215[191] | f1 |
| | Cross-model distillation | SID, ER | 231.3 h | 0.647[115] | 0.855[192] | 1.100[109] | 1.554[193] | 2.092[114] | 2.241[110] | | | | f11 |
| Code-switching | Multilingual pre-training | ASR, KWS, ST | 11.9 h | 0.342[194] 1.061[201] | 0.368[195] 1.157[202] | 0.385[196] 1.160[203] | 0.460[197] 1.293[204] | 0.512[198] 1.372[205] | 0.668[199] 1.474[56] | 0.833[66] 1.524[206] | 0.870[200] | | f1 |
| | Adapters | ST, KS, SID | 1 h | 0.510[207] 1.132[214] | 0.577[90] 1.184[215] | 0.690[208] 1.214[216] | 0.710[209] | 0.832[210] | 0.847[211] | 0.898[212] | 0.947[213] | | f2 |
| | Un-/semi-supervised learning | SLU, KWS | 154.5 h | 0.238[217] | 0.325[71] | 0.399[218] | 0.478[219] | 0.582[220] | 0.693[69] | 0.781[67] | 0.961[72] | 0.974[68] | f6 |
| Suprasegmental/ tone sensitivity | Model-layer transfer | IC, PR, LID | 11.8 h | 0.202[221] 1.266[87] | 0.236[222] 1.464[44] | 0.238[223] 1.653[46] | 0.392[47] 1.734[227] | 0.459[43] 1.742[45] | 0.818[48] 1.869[85] | 0.860[224] 1.904[228] | 1.010[225] | 1.205[226] | f1 |
| | PEFT | LID, ER, PR | 13 h | 0.168[83] 0.938[235] | 0.335[229] 0.950[77] | 0.410[230] 1.037[236] | 0.473[231] 1.068[237] | 0.664[79] 1.080[238] | 0.778[232] 1.138[82] | 0.858[233] 1.225[239] | 0.870[234] 1.408[240] | | f1 |
| | Self-supervised learning | ASR, MDD, PR | 5.5 h | 0.248[31] 1.167[245] | 0.267[241] 1.411[246] | 0.404[30] | 0.494[23] | 0.664[24] | 0.678[242] | 0.709[243] | 0.948[244] | 0.998[62] | f1 |




(a) Best single-factor associations (b) Technique-specific drivers (c) Target labeled speech requirements

Note: (a)–(b) $\Delta R^2$ from OLS with task fixed effects; dashed line = $\Delta R^2$ threshold (0.05); † = no clear technique-specific driver (top weak candidate shown, hatched bars). (c) x-axis counts target-language labeled speech only; extra unlabeled audio, multilingual source data, external text, donor-language data, and teacher-training cost are excluded.

**Figure 4.** DAG-based synthesis of the eight transfer-learning techniques for low-resource ASR. Panel (a) reports the dominant single-factor associations, panel (b) identifies technique-specific drivers after controlling for generic factors, and panel (c) summarizes practical target-label regimes, including workable thresholds and near-ideal targets.

layer transfer becomes workable at around 15 h and approaches its near-ideal regime at around 60 h. Cross-model distillation is the most label-demanding among the eight techniques, becoming workable at around 24 h and approaching its near-ideal regime at around 70 h. These values are not immutable cutoffs. They are empirical guides. They suggest that below roughly 5–7 h of target labels, multilingual pre-training, domain adaptation, PEFT, and adapters are usually the safest first choices; in the 10–15 h regime, self-supervised learning, un-/semi-supervised learning, and model-layer transfer become increasingly dependable; and once target labels reach roughly 24 h or more, distillation becomes more worthwhile, provided that the upstream teacher is already reasonably strong.

Taken together, the three panels make the application ranges of the techniques more concrete. Multilingual pre-training and self-supervised learning are best viewed as *representation-building* methods: they matter most when upstream structure is the scarce resource. Domain adaptation and un-/semi-supervised learning are *target-evidence amplifiers*: they are especially valuable when the target side has limited labels but some unlabeled in-domain speech. Distillation is a *qualitytransfer* method whose success depends less on raw scale than on whether the teacher genuinely carries better decision structure than the student. Model-layer transfer, PEFT, and adapters are *controlled parameter-reuse* mechanisms: they matter when a useful backbone already exists, but the target regime cannot safely afford full adaptation. Adapters, notably, occupy an intermediate position. They do not plateau as early as domain adaptation or semi-supervised learning, yet they are also not as purely upstream-hungry as self-supervised learning.

A further point follows from Table 4 and Figure 4: as the dominant factor increases, the gain usually begins to slow and eventually approaches a bottleneck. This pattern is unlikely to be accidental. Across all eight techniques, performance still depends, in the last analysis, on how much usable structure can be extracted from the available speech and how effectively prior acoustic knowledge can be reused in the target regime. What differs is not the existence of transfer, but which part of the learning problem each technique modifies most strongly. This observation motivates a more formal account of their common mechanism.

### *3.4 Analysis of the technology*

To formalize the empirical regularities above, we adopt the following mathematical framing as the interpretive backbone for the controlled experiments that follow.

Let the language layer be $Y \sim p_\gamma(Y)$. Let the phonological/pronunciation layer be $\Phi \sim p_\ell(\Phi \mid Y, D)$, with $D \sim p(D)$. Let the duration and alignment layer be $A \sim p(A \mid \Phi, R)$, with $R \sim p(R)$. Let the speaker and prosody/channel layer be $S \sim p(S)$, $\Pi \sim p(\Pi)$, and $U \sim p(U)$. Given the alignment $A$ and the state sequence $Q = (q_1, \ldots, q_T)$ determined by $\Phi$ and $A$, the acoustic layer is

$$p_\alpha(X \mid Q, S, \Pi, U) = \prod_{t=1}^{T} p_\alpha(x_t \mid q_t, S, \Pi, U)$$

The layers then obey the joint distribution:

$$\begin{aligned} p(X, Y, \Phi, A, S, \Pi, U, D) = p_\gamma(Y)p(D)p_\ell(\Phi \mid Y, D)p(R)p(A \mid \Phi, R) \\ p(S)p(\Pi)p(U)\prod_{t=1}^{T} p_\alpha(x_t \mid q_t, S, \Pi, U) \end{aligned} \tag{1}$$

Hence, the recognition-time posterior can be derived as:

$$p(Y \mid X) \propto \sum_{\Phi, A, S, \Pi, U, D} p_\alpha(X \mid Q, S, \Pi, U)p(A \mid \Phi, R)p_\ell(\Phi \mid Y, D)p_\gamma(Y)p(S)p(\Pi)p(U)p(D)p(R) \tag{2}$$

Here $X$ is the acoustic observation; $Y$ is the target transcription sequence; $S \in \mathcal{S}$ is the speaker; $R \in \mathcal{R}$ is speaking rate or duration; $\Pi \in \mathcal{P}$ is prosody or F0 contour; $U \in \mathcal{U}$ is noise and channel; $D \in \mathcal{D}$ is dialect or phonological variant; $\Phi = (\phi_1, \ldots, \phi_L)$ is the pronunciation sequence; and $A \in \mathcal{A}$ is the alignment from frames to phonemes or output symbols. Whether one uses a generative acoustic likelihood times language model, or an end-to-end ASR system, the basic task remains the same: integrate or sum over candidate alignments, pronunciations, and paths to obtain $p_\theta(Y \mid X)$. The differences lie in how these latent variables are modeled and approximated. The task risk under a unified test distribution $\mathcal{D}_{\text{test}}$ is then

$$R(\theta) = \mathbb{E}_{(X, Y^\star) \sim \mathcal{D}_{\text{test}}} \left[ \sum_{Y} \Delta(Y, Y^\star)\, p_\theta(Y \mid X) \right] \tag{3}$$

The No Free Lunch principle implies that no single learning algorithm can be optimal across all tasks. Good speech models therefore need task-matched bias. They must encode properties such as monotonic alignment, duration, invariance to speaker and channel variation, and reasonable text priors. Let $Q(\theta \mid \mathcal{S})$ denote a posterior and $P(\theta)$ a prior. For sample size $m$ and confidence level $\delta$, a computable upper-bound view yields

$$\mathbb{E}_{\theta \sim Q} R(\theta) \le \mathbb{E}_{\theta \sim Q} \hat{R}_S(\theta) + \sqrt{\frac{\mathrm{KL}(Q \,\|\, P) + \ln \dfrac{1}{\delta}}{2m}} + \Gamma(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}) \tag{4}$$

The three terms on the right are the empirical risk, the prior-posterior mismatch, and the domain shift.

Definition of the effective sample size $m_{\text{eff}}$: let the level $\ell \in \{\text{phn, wrd, sent, doc, mm}\}$, hours at each level $H_\ell$, unit density $r_\ell$, quality $q_\ell \in (0, 1]$, coverage $c_\ell \in (0, 1]$, and usability coefficient $\eta_\ell(M) \in [0, 1]$. Define

$$m_{\text{eff}}(D, M) = \sum_{\ell} \eta_\ell(M) r_\ell H_\ell q_\ell c_\ell \tag{5}$$

Where

$$c_\ell = \prod_{j \in \mathcal{J}_\ell} (1 - e^{-N_{\ell,j}/N^*_{\ell,j}}) \tag{6}$$

From these definitions, one obtains the empirical-risk lower bound due to incompressible noise:

$$\hat{R}_S(\theta) \gtrsim R_{\text{noise}}(D) = \sum_{\ell} w_\ell \varepsilon_\ell \tag{7}$$

where $\varepsilon_\ell$ is the annotation noise rate or equivalent noise at that level.

Definition of the domain-shift term $\Gamma$:

$$\tilde{\text{A}} = \sum_{k \in \{\text{speaker/accent, scenario, device, text domain}\}} \alpha_k \, \text{Div}_k(\text{train, test}) \tag{8}$$

Definition of the structural compatibility $\mathcal{C}_{\text{eff}}$: a structure aligned with ASR should respect monotonic alignment, local invariance, identifiability of latent variables, reasonable text priors, and limited model-task mismatch. For low-resource tonal and morphologically rich languages, it should also preserve tone contrasts, context-dependent tone alternations, and morpheme structure. This motivates the task-structure energy

$$\mathcal{E}_{\text{task}}(\theta) = \beta_{\text{inv}}\mathcal{R}_{\text{inv}}(\theta) + \beta_{\text{mon}}\mathcal{R}_{\text{mon}}(\theta) + \beta_{\text{lat}}\mathcal{R}_{\text{lat}}(\theta) + \beta_{\text{tone}}\mathcal{R}_{\text{tone}}(\theta) + \beta_{\text{morph}}\mathcal{R}_{\text{morph}}(\theta) + \beta_{\text{text}}\mathcal{R}_{\text{text}}(\gamma) + \beta_{\text{mis}}\Psi(M, D) \tag{9}$$

where:

- $\mathcal{R}_{\text{inv}}$: group-invariance regularization.
- $\mathcal{R}_{\text{mon}}$: monotonicity or alignment-consistency penalty. For CTC/RNNT, one may take

$$\mathcal{R}_{\text{mon}}(\theta) = -\ln \rho_{\text{mon}}(\theta), \quad \rho_{\text{mon}}(\theta) = \mathbb{E}_{X,Y}\left[\frac{\text{monotonic path probability mass}}{\text{total path probability mass}}\right]$$

  For AED, one may use a soft monotonicity violation measure

$$\mathcal{R}_{\text{mon}}(\theta) = \mathbb{E}_X\left[\sum_n \sum_{t' < t} \alpha_{n,t'}\alpha_{n+1,t}\right]$$

- $\mathcal{R}_{\text{lat}}$: identifiability of latent variables, measured for example by the optimal small-capacity classification error $\varepsilon_j(\theta)$ on probe tasks,

$$\mathcal{R}_{\text{lat}}(\theta) = \sum_j \lambda_j \mathcal{E}_j(\theta)$$

- $\mathcal{R}_{\text{tone}}$ : tonal discriminability and sandhi consistency. A generic form is

$$\mathcal{R}_{\text{tone}}(\theta) = \lambda_{\text{pair}} \mathbb{E}\Big[\max\big(0,\ m - d(z_i,\ z_j)\big)\Big] + \lambda_{\text{sd}} \mathbb{E}\Big[\ell\big(\hat{\tau}_{1:n},\ \mathcal{S}(\tau_{1:n},\ c)\big)\Big]$$

  where $z_i$, $z_j$ are embeddings of tone-minimal-pair syllables, $\tau$ denotes lexical tones, and $\mathcal{S}$ applies a context-dependent sandhi rule under context $c$.

- $\mathcal{R}_{\text{morph}}$ : morpheme-boundary and morpho-syntactic feature consistency, for example

$$\mathcal{R}_{\text{morph}}(\theta) = \mathbb{E}\Big[\ell_{\text{seg}}(\hat{b},\ b) + \ell_{\text{feat}}(\hat{\mu},\ \mu)\Big]$$

  where $b$ is the morpheme-boundary sequence and $\mu$ is a morphological feature bundle such as case, tense, number, or voice.

- $\mathcal{R}_{\text{text}}(\gamma) = \ln(\text{ppl}_\gamma / \text{ppl}^\star)_+$ : text-prior consistency.
- $\Psi(M,\ D) \geq 0$: structural mismatch penalty. One example is a task that requires online monotonicity while the model still uses unrestricted global attention.

Lower-energy parameters are more probable under the prior, so together with a base prior $P_0$ we obtain the task-structure prior

$$P_{\text{task}}(\theta) \propto \exp(-\mathcal{E}_{\text{task}}(\theta)) P_0(\theta) \tag{10}$$

where $P_0$ may be an isotropic Gaussian or another generic base prior.

Therefore, the structural compatibility can be defined as

$$\mathcal{C}_{\text{eff}}(M,\ D) \triangleq \mathbb{E}_S\left[\text{KL}(Q_M(\cdot \mid \mathcal{S}) \,\|\, P_{\text{task}})\right] \tag{11}$$

The smaller the KL to the task-structure prior, the more the posterior adheres to the relevant ASR structure. Thus the two terms of the upper bound can be approximately estimated as

$$\mathcal{C}_{\text{eff}}(M,\ D) \lesssim \underbrace{\mathbb{E}_S \text{KL}(Q_M \,\|\, P_0)}_{\text{Information/parameter complexity}} + \underbrace{\mathbb{E}_{\theta \sim Q_M} \mathcal{E}_{\text{task}}(\theta)}_{\text{Consistency with task structure}} \tag{12}$$

Substituting $m_{\text{eff}}$, $\Gamma$, and $\mathcal{C}_{\text{eff}}$ back into Eq. (4), we obtain the unified ASR operational risk upper bound

$$\mathbb{E}_{\theta \sim Q_M} R(\theta) \leq \underbrace{\mathbb{E}_{\theta \sim Q_M} \hat{R}_S(\theta)}_{\text{Empirical risk}} + \underbrace{\sqrt{\frac{C_{\text{eff}}(M,\ D) + \ln\frac{1}{\delta}}{2m_{\text{eff}}(D,\ M)}}}_{\text{Complexity-data trade-off term}} + \underbrace{\Gamma(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})}_{\text{Domain shift term}} \tag{13}$$

This bound clarifies the roles of the main transfer techniques. Self-supervised learning primarily reduces operational risk by lowering $\mathcal{C}_{\text{eff}}$ and partly suppressing $\Gamma$, rather than by directly increasing $m_{\text{eff}}$. It does so by learning representations that are less sensitive to irrelevant variation in $S$, $\Pi$, $U$, $R$, which lowers $\mathcal{R}_{\text{inv}}$. Multilingual learning acts through both $\mathcal{C}_{\text{eff}}$ and $m_{\text{eff}}$: it shares parameters and output structure across languages, while broadening the coverage term $c_\ell$. The same logic also explains why language-specific inductive bias should not be relegated to black-box scale. Tone-aware constraints reduce $\mathcal{R}_{\text{tone}}$ and protect contrasts that generic invariance objectives may smooth away. Morphology-aware output modeling reduces $\mathcal{R}_{\text{morph}}$ by aligning the decoder with stem-affix composition rather than whole-word sparsity. In both cases, the primary gain appears as a reduction in $\mathcal{C}_{\text{eff}}$, and the secondary gain appears as an increase in usable supervision because each labeled utterance yields more linguistically organized constraints.

The eight dominant factors extracted by DAG can also be read through Eq. (13). Multilingual pretraining is mainly governed by $f_1$ and $f_7$, because upstream hours and language coverage together enlarge coverage and reduce structural burden. Self-supervised learning is more tightly tied to $f_1$, because more upstream unlabeled speech mainly lowers $\mathcal{R}_{\text{inv}}$ and therefore $\mathcal{C}_{\text{eff}}$. Cross-model distillation is governed by $f_{11}$, because teacher quality directly changes the surrogate supervision seen by the student. Domain adaptation is naturally controlled by $f_6$, since additional unlabeled target speech sharpens the estimate of the train-test divergence and acts directly on $\Gamma$. Un-/semi-supervised learning depends on both $f_6$ and $f_9$: more unlabeled data can enlarge $m_{\text{eff}}$, but only if the labeled-to-unlabeled ratio remains sufficient to keep pseudo-label noise under control. Model-layer transfer is explained by the combination of $f_1$ and $f_4$. PEFT follows the same logic, with $f_1$ reflecting prior quality and $f_{10}$ the size of the adaptation subspace. Adapters are especially sensitive to $f_2$ and $f_{10}$, because their benefit depends on whether backbone capacity and bottleneck width are jointly sufficient to express the target-specific correction.

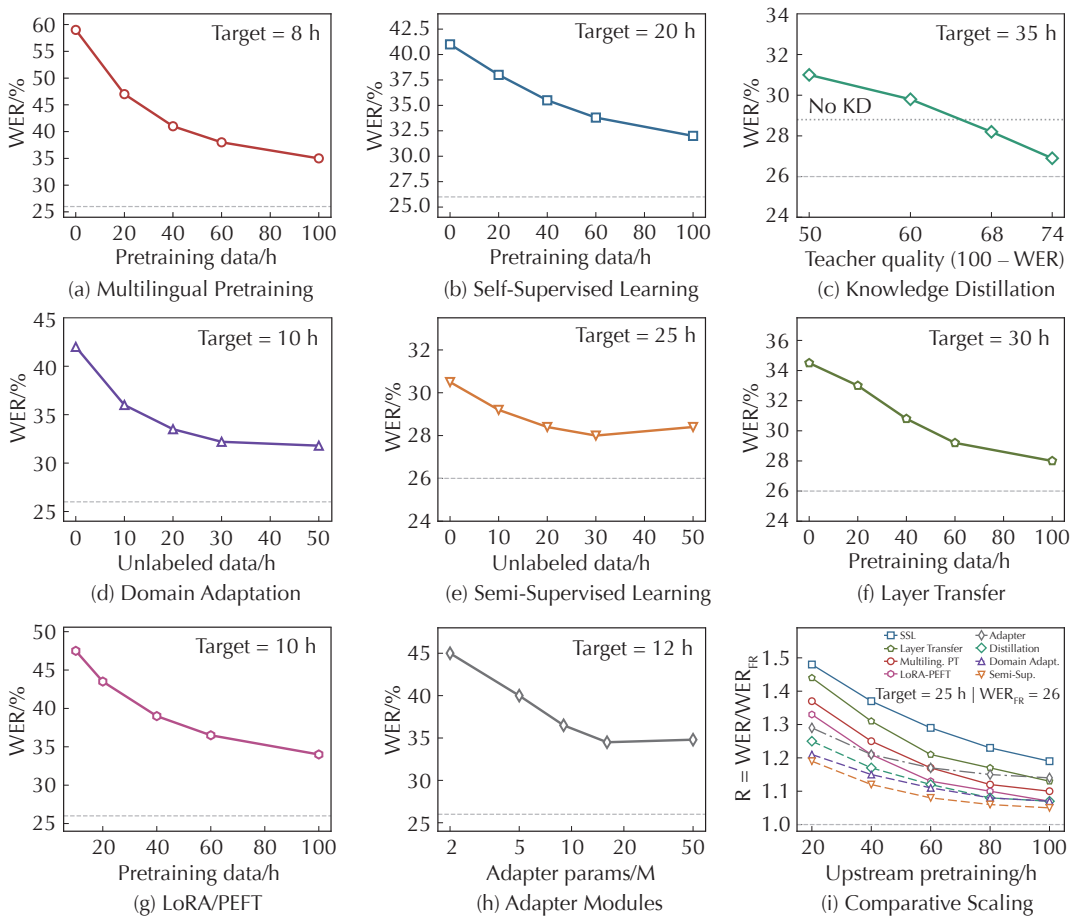
### *3.5 Controlled validation under unified settings*

To examine whether this term-wise interpretation survives under controlled conditions, we ran a unified experiment in which each technique is forced to vary primarily along its theorized main factor while the rest of the setup is kept deliberately simple.

- **Data.** The main corpus is Common Voice 25.0, with Italian as the target language and French, German, Spanish, and Portuguese as source languages.
- **Reference point.** The full-resource baseline is a 200 h Italian system with a 16M Conformer-CTC backbone, no external language model, and test WER $\approx$ 26%.
- **Unified backbone.** All eight techniques are implemented on the same lightweight backbone family: 80-dim log-Mel input, greedy CTC decoding, no external LM. SSL uses a lightweight masked acoustic modeling variant rather than heavy raw-waveform pretraining; PEFT uses LoRA; adapters use bottleneck modules.

- **Protocol.** The first eight plots sweep only the factor identified as dominant by the DAG analysis: upstream labeled hours for multilingual pre-training, model-layer transfer, and PEFT; upstream unlabeled hours for SSL; teacher quality for distillation; target unlabeled hours for domain adaptation and un-/semi-supervised learning; and backbone size for adapters. For methods other than the pretraining-hours family, the upstream condition is fixed near a common operating point. The ninth plot fixes target labeled data at 25 h for all techniques and compares them over 20/40/60/80/100 h of upstream scale via $R = \mathrm{WER}_{\mathrm{method}}/\mathrm{WER}_{\mathrm{FR}}$.

The results are presented in Figure 5.




Dashed gray line in (a)-(h): full-resource baseline (WER = 26). In (i): solid = pretraining-scale dependent; dashed = mechanism-design dependent; dash-dot = intermediate.

**Figure 5.** Controlled validation under unified experimental settings. Panels (a)–(h) vary the dominant factor for each of the eight transfer-learning techniques, while panel (i) compares all methods on the common normalized scale $R = \mathrm{WER}_{\mathrm{method}}/\mathrm{WER}_{\mathrm{FR}}$.

The unified ninth panel uses 25 h of labeled Italian as a common comparison point. This choice is an engineering compromise: it lies above the workable threshold of the most label-demanding techniques, but is already comparatively generous for the most label-efficient ones. The absolute ranking in that panel should therefore be read with caution; the more informative signal is the *shape* of each response curve and how it changes with the dominant factor.

*Multilingual pre-training*. With 8 h of labeled target speech, WER decreases from 59 → 47 → 41 → 38 → 35 as upstream labeled hours increase from 0 to 100. In the unified ninth panel, the ratio improves from $R = 1.37$ at 20 h to 1.10 at 100 h. This steady but gradually saturating decline is exactly what one would expect if the main benefit comes from enlarging shared cross-lingual coverage and lowering effective complexity. Its bottleneck is therefore not adaptation capacity per se, but the diminishing marginal value of additional upstream diversity once the remaining mismatch becomes increasingly target-specific.

*Self-supervised learning*. With 20 h of target labels, WER falls more slowly, from 41 → 38 → 35.5 → 33.8 → 32.0, and the unified ratio improves from 1.48 to 1.19. The gain is real, but more upstream-hungry than multilingual pre-training. This supports the theoretical picture that SSL first reduces representational complexity and invariance burden, and only later converts that advantage into downstream accuracy. Its bottleneck is that better generic speech representations do not automatically resolve all target-language or task-specific ambiguities.

*Cross-model distillation*. Distillation is the clearest case in which a weak mechanism can be harmful. At 35 h of target labels, the no-KD system starts at 28.8 WER, but a weak teacher with dev WER ≈ 50 degrades the student to 31.0. Only as teacher quality improves does the student recover and then benefit: 29.8, 28.2, and finally 26.9 when the teacher dev WER reaches 26. The unified ratio correspondingly improves from 1.25 to 1.07. This strongly supports the idea that distillation operates through the quality of the transferred decision structure: a strong teacher lowers the student's effective empirical risk, whereas a weak teacher injects structured noise.

*Domain adaptation*. At 10 h of labeled target speech, increasing target-side unlabeled data from 0 to 50 h reduces WER from 42 → 36 → 33.5 → 32.2 → 31.8. In the unified panel, the ratio improves from 1.21 to 1.07. The curve is steep early and then flattens quickly, which is exactly the signature of a method that primarily attacks $\Gamma(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$: once the most salient mismatch has been aligned, additional unlabeled data yields smaller returns.

*Un-/semi-supervised learning*. At 25 h of target labels, WER improves from 30.5 → 29.2 → 28.4 → 28.0 as unlabeled target hours increase, but then slightly rebounds to 28.4 at the largest unlabeled budget. In the unified panel, the method is the strongest at 100 h with $R \approx 1.05$. The improvement confirms that semi-supervised learning can enlarge $m_{\text{eff}}(D, M)$, yet the mild rebound is equally informative: once pseudo-label noise begins to accumulate faster than usable supervision, the gain saturates and may reverse. Its bottleneck is therefore not unlabeled scale alone, but the stability of the pseudo-labeling process.

*Model-layer transfer.* At 30 h of target labels, WER decreases monotonically but modestly, 34.5 → 33.0 → 30.8 → 29.2 → 28.0, and the unified ratio moves from 1.44 to 1.13. The curve is cleaner than it is dramatic. This is consistent with the theoretical role of layer transfer: selective freezing stabilizes optimization and preserves a useful prior, but it cannot fully escape the representational ceiling inherited from upstream training. The bottleneck lies in the balance between retained prior structure and the limited plasticity of the layers that remain trainable.

*LoRA-style PEFT.* At 10 h of target labels, WER drops from 47.5 → 43.5 → 39.0 → 36.5 → 34.0, while the unified ratio improves from 1.33 to 1.07. The response curve is both clean and practically attractive: a stronger upstream prior continues to help, yet the method can already convert moderate upstream scale into strong downstream gains under a tightly constrained trainable budget. The bottleneck is the joint ceiling imposed by backbone quality and low-rank adaptation capacity. Once either of them becomes limiting, additional upstream scale translates into smaller downstream improvement.

*Adapters.* At 12 h of target labels, increasing backbone size from 2M to 16M lowers WER from 45.0 → 40.0 → 36.5 → 34.5, but the largest 50M backbone yields 34.8 rather than another improvement. In the unified panel, the method ends at $R \approx 1.14$. This is a particularly instructive result: larger backbones help only up to the point at which bottleneck width, optimization stability, and fixed target supervision cease to support the extra capacity. Adapters therefore sit between scale-driven and mechanism-driven techniques. They benefit from a stronger backbone, but not monotonically or indefinitely.

Taken together, the nine panels in Figure 5 provide a controlled empirical check of the mathematical picture in Eq. (13). Whenever the factor that mainly governs $\mathcal{C}_{\text{eff}}$, $m_{\text{eff}}$, $\Gamma$, or teacher-induced noise is changed in the expected direction, WER changes in the expected direction as well. In this sense, the experiments support the term-wise interpretation of the bound. They also clarify the bottleneck structure of the field. Multilingual pre-training, self-supervised learning, model-layer transfer, and LoRA-PEFT are visibly more dependent on upstream scale; domain adaptation, un-/semi-supervised learning, and distillation derive more of their benefit from mechanism design and enter their plateaus earlier; adapters occupy a middle position. The common pattern, however, is saturation. No method keeps gaining linearly, and none escapes the underlying limits imposed by data quality, data quantity, and the finite reusability of inherited acoustic knowledge.

## 4. FUTURE DIRECTIONS

### *4.1 From efficient reuse to richer learning*

The analysis above suggests a structural conclusion. Sufficiently large and sufficiently clean corpora remain the most fundamental resource in low-resource ASR, and building them is usually slow. While such corpora are being built, the key question is how to obtain stronger training signals from the available speech and how to incorporate prior knowledge while minimizing mismatch with the target setting. The eight mainstream techniques can be read as different answers to that question.

At a broad level, they fall into two intuitive families. The first is *corpus-efficiency oriented*: multilingual pre-training, self-supervised learning, domain adaptation, and un-/semi-supervised learning attempt to increase the usable information content of available speech or to align it better with the target setting. The second is *parameter-knowledge-transfer oriented*: cross-model distillation, model-layer transfer, PEFT, and adapters attempt to inject, preserve, or route knowledge already stored in other models or pretrained weights. The distinction matters because it explains why their bottlenecks differ. Corpus-efficiency methods eventually saturate because additional data or unlabeled audio yields diminishing structural novelty and may accumulate noise. Parameter-transfer methods saturate because no adaptation strategy can consistently exceed the ceiling set by the inherited prior and the limited target evidence.

This is why the present plateau is unlikely to be overcome by simply turning any single existing knob harder. The eight mainstream techniques are, in different ways, methods for *using existing information more efficiently*. What they do not fundamentally change is the *order* of information available to the learner. Once that limit is reached, methods that reorganize supervision itself become attractive.

Two such directions are especially promising. The first is **multi-task learning**. It jointly captures phonetic, lexical, prosodic, speaker, semantic, and sometimes multimodal regularities, thereby converting more latent structure into usable supervision and reducing structural mismatch at several levels simultaneously. The second is **meta-learning**. It learns better priors and initialization rules across task families, so that adaptation depends not only on the amount of data but also on recurring task structure. At the current frontier, **large-scale multi-task speech models** combine both ideas. They absorb massive multilingual and multi-domain corpora, encode knowledge in large parameter spaces, and can be paired with efficient adaptation modules. This does not remove the need for better corpora. It offers a stronger way to turn heterogeneous resources into transferable structure.

The mechanism-oriented synthesis in Table 5 serves as the conceptual bridge for the discussion that follows.

**Table 5.** Systematic Analysis of Technical Principles.

| Phenomenon/ resource | Mechanism and influencing variables (condensed) | Contribution to Eq. (13) | Typical effects |
|---|---|---|---|
| Phoneme-levelannotations | Constrain $p_\ell(\Phi \mid Y, D)$ and $p(A \mid \Phi, R)$; reduce $\mathcal{R}_{\text{mon}}$ and $\mathcal{R}_{\text{lat}}$; increase $\eta_{\text{phn}}$ | $m_{\text{eff}} \uparrow$, $\mathcal{C}_{\text{eff}} \downarrow$, $R_{\text{noise}} \downarrow$ | Stable alignment; faster convergence |
| Character/word transcription | Directly supervise $Y$; increase $\eta_{\text{wrd}}$ | $m_{\text{eff}} \uparrow$, $R_{\text{noise}} \downarrow$ | Lower NLL/MBR |
| Tone/F0 supervision | Tone embeddings, F0-conditioned layers, tone heads, or sandhi constraints preserve lexical pitch contrasts and contextual tonal alternations; reduce $\mathcal{R}_{\text{tone}}$ and part of $\mathcal{R}_{\text{lat}}$ | $\mathcal{C}_{\text{eff}} \downarrow$, $\Gamma \downarrow$ | Better tone minimal-pair separation; fewer sandhi errors |

**Table 5.** *Continued.*

| Phenomenon/ resource | Mechanism and influencing variables (condensed) | Contribution to Eq. (13) | Typical effects |
|---|---|---|---|
| Morphology-aware output modeling | Linguistically guided segmentation, stem-affix factorization, and inflectional-tag heads reduce output sparsity and lower $\mathcal{R}_{\text{morph}}$ | $m_{\text{eff}} \uparrow$, $\mathcal{C}_{\text{eff}} \downarrow$ | Better rare-form generalization; fewer OOVs |
| Sentence-/ document-level supervision | Strengthen $p_\gamma(Y)$; reduce $\mathcal{R}_{\text{text}}(\gamma)$; improve long-range consistency | $m_{\text{eff}} \uparrow$, $\mathcal{C}_{\text{eff}} \downarrow$, $\Gamma \downarrow$ | Better coherence; constrained decoding |
| Multimodal alignment | Cross-modal semantic constraints reduce $\mathcal{R}_{\text{lat}}$ and help bridge domain mismatch | $m_{\text{eff}} \uparrow$, $\mathcal{C}_{\text{eff}} \downarrow$, $\Gamma \downarrow$ | Semantic disambiguation; noise robustness |
| Unlabeled speech (SSL) | Invariance priors reduce $\mathcal{R}_{\text{inv}}$ and shape transferable acoustic structure | $\mathcal{C}_{\text{eff}} \downarrow$ | Lower complexity; faster adaptation |
| Multilingual sharing | Shared parameters, phonology, and subwords reduce separate per-language complexity and improve coverage | $\mathcal{C}_{\text{eff}} \downarrow$, $\Gamma \downarrow$, $m_{\text{eff}} \uparrow$ | Better transfer with fewer labels |
| External/internal LM | Strengthens $p_\gamma(Y)$; reduces $\mathcal{R}_{\text{text}}(\gamma)$ and textual domain mismatch | $\mathcal{C}_{\text{eff}} \downarrow$, $\Gamma \downarrow$ | Better decoding under domain shift |
| Data augmentation/ invariance | Approximate group invariance through perturbation; improve robustness to noise, speaker, channel, and speed variation | $\mathcal{C}_{\text{eff}} \downarrow$, $\Gamma \downarrow$ | Stronger generalization |
| *The following three approaches jointly optimize multiple terms in Eq. (13) and therefore have broader long-term potential* | | | |
| Multi-task learning | Jointly constrain latent factors $S$, $\Pi$, $D$, $\Phi$, $A$, and optionally tone and morphology; reduce $\mathcal{R}_{\text{lat}}$, $\mathcal{R}_{\text{mon}}$, $\mathcal{R}_{\text{inv}}$, $\mathcal{R}_{\text{tone}}$, $\mathcal{R}_{\text{morph}}$ while adding auxiliary supervision | $\mathcal{C}_{\text{eff}} \downarrow$, $m_{\text{eff}} \uparrow$, $\Gamma \downarrow$ | Shared transferable representations; strong low-resource gains |
| Meta-learning/good initialization | Learn task-conditioned priors and update rules across tasks; directly reduce the distance from initialization to target optimum and improve few-shot adaptation geometry | $\mathcal{C}_{\text{eff}} \downarrow$, $\Gamma \downarrow$ | Fastest few-shot convergence; better rapid adaptation |
| Large-scale multi-task speech models | Unify $p_\ell$, $p_\alpha$, $p_\gamma$ in a shared foundation model; broad pretraining simultaneously reduces invariance, alignment, text, tone, and morphology penalties while enlarging coverage | **all three :** $m_{\text{eff}} \uparrow$, $\mathcal{C}_{\text{eff}} \downarrow$, $\Gamma \downarrow$ | Best overall risk reduction; emergent cross-task generalization |

The following therefore focuses on three themes in turn: new transfer-learning paradigms that exploit structure more directly, corpus expansion and information extraction strategies that raise the usable supervision of speech, and large-scale multi-task speech models that integrate both data scale and task knowledge within a single framework.

### *4.2 Task-Based Transfer Learning Paradigm*

Task-based transfer is the first place where the theoretical tools developed above become fully operational rather than merely descriptive. The preceding discussion showed that low-resource ASR is constrained by empirical noise, the complexity-data trade-off, and domain shift. It also showed that many strong current methods are still largely driven by scale-related factors and eventually saturate once upstream diversity is absorbed. Meta-learning and multi-task learning become attractive precisely at that point. They do not simply add more hours. They reorganize supervision at the level of tasks, episodes, and auxiliary constraints.

To make this connection explicit, consider a fixed operating point $(\hat{R},\ \mathcal{C}_{\text{eff}},\ m_{\text{eff}},\ \Gamma)$ and let $\Delta$ denote the change induced by a task-based method relative to a corpus-only baseline with the same backbone and decoding setup. A first-order expansion of Eq. (13) yields

$$\Delta\mathfrak{B}_{\text{task}} \approx \Delta\hat{R} + \frac{\Delta\mathcal{C}_{\text{eff}}}{\sqrt{8m_{\text{eff}}\mathcal{C}_{\text{eff}}}} - \frac{\sqrt{\mathcal{C}_{\text{eff}}}}{\sqrt{8}m_{\text{eff}}^{3/2}}\Delta m_{\text{eff}} + \Delta\Gamma \tag{14}$$

where $\mathfrak{B}_{\text{task}}$ denotes the task-specific slack of the bound. This expression clarifies the difference between the two families. Meta-learning acts mainly through $\Delta\mathcal{C}_{\text{eff}} < 0$, and often also through $\Delta\Gamma < 0$, because it learns how to update under task shift. Multi-task learning can act on the same two terms, but it has an additional route: it can make $\Delta m_{\text{eff}} > 0$ by turning one utterance into several structured constraints.

#### *4.2.1 Meta Learning*

*Theoretical advantages.* Meta-learning is best understood as prior shaping in update space. The main object is not a representation alone, but the distance that must still be traveled from a shared initialization to the optimum of a new task. Let $\tau$ denote a task, let $\theta_\tau^\star = \arg\min_\theta L_\tau(\theta)$, and let $P_{\theta_0,\ \sigma^2 I}$ be a Gaussian prior centered at initialization $\theta_0$. Under a local quadratic approximation with task posterior $Q_\tau \approx \mathcal{N}(\theta_\tau^\star,\ H_\tau^{-1})$, the task-conditioned contribution to the complexity term is

$$\mathrm{KL}(Q_\tau \,\|\, P_{\theta_0,\ \sigma^2 I}) \approx \frac{d_\tau(\theta_0)}{2\sigma^2} + \frac{1}{2}[\sigma^{-2}\mathrm{tr}(H_\tau^{-1}) - \ln|H_\tau^{-1}| + c_0], \quad d_\tau(\theta_0) = \|\theta_\tau^\star - \theta_0\|_2^2 \tag{15}$$

where $H_\tau$ is the local Hessian and $c_0$ does not depend on $\theta_0$. Meta-learning is attractive because it directly shrinks the average task-conditioned displacement $d_\tau(\theta_0)$.

The meta objective makes this geometry explicit. Writing $S_\tau$ and $Q_\tau$ for support and query sets, and $U_K$ for $K$ inner-loop updates, we have

$$\theta_{\text{meta}} = \arg\min_{\theta_0} \mathbb{E}_{\tau\sim p(\tau)} L_\tau^Q\left(U_K(\theta_0;\, S_\tau)\right) \tag{16}$$

For one inner step with learning rate $\alpha$, the query loss admits the expansion

$$L_\tau^Q(\theta_0 - \alpha g_\tau^S) \approx L_\tau^Q(\theta_0) - \alpha\langle g_\tau^Q,\, g_\tau^S\rangle + \frac{\alpha^2}{2}(g_\tau^S)^\top H_\tau^Q g_\tau^S \tag{17}$$

where $g_\tau^S = \nabla_\theta L_\tau^S(\theta_0)$ and $g_\tau^Q = \nabla_\theta L_\tau^Q(\theta_0)$. Meta-learning therefore favors initializations for which support gradients are predictive of query improvement.

Under the same local approximation, one may read meta-learning as minimizing average adaptation radius:

$$\mathbb{E}_\tau d_\tau(\theta_{\text{meta}}) \lesssim \mathbb{E}_\tau d_\tau(\theta_{\text{pt}}) \tag{18}$$

for any corpus-only initialization $\theta_{\text{pt}}$ in the same hypothesis family. The point is not that this statement is distribution-free. It is that meta-learning and corpus-driven pretraining solve related but distinct problems. Corpus-driven transfer improves what the model already knows. Meta-learning improves how efficiently the model can move toward what the new task requires.

*Application methodology.* The strongest meta-learning results are those that align inner-loop evidence, outer-loop evaluation, and source-task transfer directions with the actual bottleneck of the target problem. In code-switching and multilingual ASR, methods such as Two-Fold CV and language-as-task training reduce the mismatch between the tasks seen in the inner and outer loops [247-249]. In the language of Eq. (17), they increase the chance that the inner update remains query-predictive.

A second family of studies addresses task imbalance. Easy or high-resource tasks can monopolize the meta-gradient. Difficulty-aware reweighting, adversarial meta-sampling, and task-level polynomial meta-losses are all responses to this problem [250-252]. Their common form can be summarized as

$$\min_{\theta_0,\,\mathbf{w}} \sum_\tau w_\tau L_\tau^Q\left(U_K(\theta_0;\, S_\tau)\right), \quad n_{\text{eff}}(\mathbf{w}) = \frac{1}{\sum_\tau w_\tau^2} \tag{19}$$

where $n_{\text{eff}}(\mathbf{w})$ is the effective number of tasks that meaningfully shape the outer objective. Once this quantity collapses, meta-learning degenerates into repeated adaptation to the same easy region.

A third family makes the update geometry itself more target-tangent. Projection-corrected meta-transfer and related gradient filtering methods exemplify this logic [247, 253]. Their shared idea can be written as

$$g_{\text{src}\to\tau} = g_{\text{src}} - \lambda\Pi_{\perp g_\tau}(g_{\text{src}}) \tag{20}$$

which removes the component of the source-task gradient that is orthogonal to the target descent direction. This is a direct mechanism for reducing structural mismatch.

Meta-learning can also enlarge the usable support of scarce data even when raw hours do not change. The gain comes from episode construction rather than from more acoustic evidence. If $\kappa_{\ell,e} \in [0, 1]$ denotes the probability that episode $e$ contributes a new usable constraint at level $\ell$, one may write

$$m_{\text{eff}}^{\text{epi}} = \sum_{\ell} r_{\ell} H_{\ell} q_{\ell} \left[ 1 - \prod_{e=1}^{E} (1 - \kappa_{\ell,e}) \right] \tag{21}$$

This explains why task augmentation in child speech, support/query reshuffling in code-switching, and family-aware task sampling can help without changing the nominal corpus size [254-255].

*Language-specific episodic design.* Meta-learning becomes more compelling when episodes are built from the structure that generic pretraining usually under-models. For tonal languages, tasks can be defined around tone-minimal pairs, speaker-normalized F0 families, or sandhi contexts. For morphologically rich languages, tasks can be organized around inflectional paradigms, stem reuse, or case/tense/voice alternations. A generic episode loss can be written as

$$L_{\tau} = L_{\text{ASR}}^{(\tau)} + \lambda_{\text{tone}} L_{\text{tone}}^{(\tau)} + \lambda_{\text{morph}} L_{\text{morph}}^{(\tau)}$$

with

$$L_{\text{tone}}^{(\tau)} = L_{\text{cls}}(\hat{\tau}, \tau) + \eta L_{\text{sandhi}}, \quad L_{\text{morph}}^{(\tau)} = L_{\text{seg}}(\hat{b}, b) + L_{\text{feat}}(\hat{\mu}, \mu)$$

This formulation is useful because it shows how language-specific priors enter the update rule rather than only the final decoder. A tone episode asks whether the model can adapt quickly while preserving tone contrasts. A morphology episode asks whether it can generalize across paradigms rather than memorize isolated forms. These ideas are consistent with work on unified alphabets, cross-script units, prompt-tuned audio-language models, and rapid transfer under accent or task shift [256-261].

*Additional notes.* Several variants show that the benefit of meta-learning survives substantial changes in architecture and label space. Metric-based few-shot recognition such as MN-CTC reduces the admissible hypothesis class rather than explicitly learning an optimizer [262]:

$$\mathfrak{R}_n(\mathcal{H}_{\text{metric}}) \leq \mathfrak{R}_n(\mathcal{H}_{\text{full}}) \tag{22}$$

where $\mathfrak{R}_n$ is the empirical Rademacher complexity. Unit redesign provides a second example. Work on unified alphabets, components, and cross-script sublexical units shows that meta-learning becomes stronger when the label space itself carries cross-task regularity [256-257]. This effect can be expressed as a reduction in pronunciation uncertainty:

$$\Delta_{\text{unit}} = H(\Phi \mid Y, D)_{\text{old}} - H(\Phi \mid Y, D)_{\text{new}} > 0 \tag{23}$$

Prompt-based adaptation for large audio–language models reaches the same conclusion from another angle [258]. Let $z \in \mathbb{R}^s$ be a deep prompt with $s \ll d = \dim(\theta)$, and let $\kappa_J$ summarize the conditioning of the prompt-to-parameter Jacobian. Then a low-dimensional adaptation subspace gives

$$\mathrm{KL}(Q_z \,\|\, P_z) \le \kappa_J \frac{S}{d} \mathrm{KL}(Q_\theta \,\|\, P_\theta) \tag{24}$$

Accent adaptation, Conformer-Transducer meta-transfer, and ASR + NLU task-level adaptation support the same general reading [259, 263-264].

*A practical reading.* Meta-learning is most attractive when the residual difficulty lies in rapid adaptation rather than in the complete absence of upstream knowledge. It is especially valuable for languages where the decisive structure is highly reusable but sparsely observed, such as tone classes, sandhi contexts, or inflectional paradigms. In the notation of Eq. (13), this is the regime in which reducing $\mathcal{C}_{\text{eff}}$ through a better task-conditioned prior matters more than simply adding more raw hours [265].

### *4.2.2 Multi-task Learning*

*Core idea.* Multi-task learning is most useful when the transcript alone does not fully specify which latent factors should be preserved, ignored, or disentangled. The basic object is a joint objective

$$L_{\text{MTL}}(\theta) = L_0(\theta) + \sum_{a=1}^{A} \lambda_a L_a(\theta), \quad g_{\text{MTL}} = g_0 + \sum_{a=1}^{A} \lambda_a g_a \tag{25}$$

where $L_0$ is the target ASR loss and $L_a$ are auxiliary tasks. A single update with step size $\eta$ changes the target loss according to

$$L_0(\theta - \eta g_{\text{MTL}}) \approx L_0(\theta) - \eta \,\|\, g_0 \,\|_2^2 - \eta \sum_a \lambda_a \langle g_0,\, g_a \rangle + \frac{\eta^2}{2} g_{\text{MTL}}^{\top} H_0 g_{\text{MTL}} \tag{26}$$

where $H_0$ is the Hessian of the target loss. Auxiliary tasks help only if their gradients align with the target descent or regularize it through curvature.

This can be summarized by a task-alignment matrix

$$A_{ij} = \frac{\langle g_i,\, g_j \rangle}{\| g_i \|_2 \| g_j \|_2}, \quad \mathbf{A} \in [-1,\, 1]^{(A+1)\times(A+1)} \tag{27}$$

This is why gating, experts, language IDs, local attention, and curriculum schedules are not merely engineering details. They are ways of reshaping **A** so that positive transfer occupies shared parameters and negative transfer is routed away.

Multi-task learning also differs from simple data scaling because it can enlarge the equivalent supervision of each utterance. Let $\mathcal{A}_\ell$ denote the set of auxiliary heads that constrain level $\ell$. Then a multi-view effective sample size can be written as

$$m_{\text{eff}}^{\text{MTL}} = \sum_{\ell} r_\ell H_\ell q_\ell \left( 1 - \prod_{a \in \mathcal{A}_\ell} (1 - \eta_{\ell,a} c_{\ell,a}) \right) \tag{28}$$

where $\eta_{\ell,a}$ is the usability of auxiliary task $\alpha$ for level $\ell$, and $c_{\ell,a}$ is its coverage at that level.

*Application domains.*

**Label-space** level Label-space MTL is the clearest demonstration of the theory because it directly attacks phonological ambiguity and alignment sparsity. Mixed-granularity systems that parallelize CD/CI targets, phonemes, syllables, senones, or native and universal grapheme/phone sets all work by adding extra views of the same latent path [266-267]. A successful design enforces

$$H(\Phi,\ A \mid Z,\ Y)_{\text{MTL}} < H(\Phi,\ A \mid Z,\ Y)_{\text{STL}} \tag{29}$$

where $Z$ is the learned representation.

Joint supervised-unsupervised formulations extend the same logic to unlabeled speech. JUST and JSUM are representative examples [268-269]. They are useful when the unsupervised objective injects additional constraints without rotating the representation away from ASR. In terms of Eq. (26), the critical condition is that $\langle g_0, g_a \rangle$ remains non-negative around the operating point.

**Handling speech variability** When the main bottleneck is variability rather than label granularity, the goal of MTL changes. The model now tries to separate nuisance factors from the transcription path. A generic formulation is

$$\min_{\theta}\ L_{\text{ASR}}(\theta) + \sum_{v} \lambda_v L_v(\theta) + \mu \sum_{v} I(Z_{\text{inv}};\ V_v) \quad \text{s.t.} \quad I(Z_{\text{inv}};\ Y) \geq \kappa \tag{30}$$

where $V_v$ may denote accent, dialect, speaker, genre, or modality, and $Z = [Z_{\text{inv}},\ Z_{\text{spec}}]$ is understood as an invariant ASR-facing subspace together with task-specific channels. Mixtures of informed experts, multilingual CTC/attention with language IDs, multi-domain adaptation, dialect-aware auxiliary heads, mid-layer cross-attention injection, and local-attention WaveNet-CTC systems all implement versions of this principle [270-276].

**Tone and morphology as auxiliary structure** Tone and morphology are especially suitable auxiliary targets because they expose structure that a transcript alone often leaves implicit. A practical objective is

$$L = L_{\text{ASR}} + \lambda_{\text{F0}} L_{\text{F0}} + \lambda_{\text{tone}} L_{\text{tone}} + \lambda_{\text{sandhi}} L_{\text{sandhi}} + \lambda_{\text{seg}} L_{\text{morph-seg}} + \lambda_{\text{tag}} L_{\text{morph-tag}}$$

Tone heads encourage the shared encoder to preserve F0 trajectories and lexical tone contrasts. Sandhi heads make contextual tonal alternations explicit. Morphology heads encourage the decoder to factor rare words into reusable stems and affixes. Compared with plain BPE, linguistically guided segmentation often sacrifices some tokenizer simplicity, but it can yield better transfer in agglutinative settings because updates are reused across many inflected forms [49-50, 52, 54]. These losses are also orthogonal to SSL and PEFT: SSL supplies a strong acoustic prior, PEFT controls update cost, and the tone or morphology heads determine which distinctions the transfer pipeline must preserve [23, 43-44, 91-92].

**SLU** SLU-oriented **MTL** is valuable because it turns downstream semantic, retrieval, and enhancement signals into extra supervision for the acoustic representation. Work on joint phoneme recognition, keyword discovery, topic or intent prediction, query-by-example spoken term detection, and multi-example KWS supports this point from several angles [277-279]. Noise-robust enhancement and affective or intent-related coupling provide an informative stress test. Enhancement optimized specifically for ASR reduces the effective noise floor because the front end is not asked to denoise in the abstract, but to do so in a way that preserves recognizability [280]. Cross-lingual shared hidden layers and ASR-pretrained encoders for intent or emotion recognition also help when the shared subspace captures transferable acoustic structure [281-282]. Negative cases remain important. Under strong interfering speech, ASR and SER may compete rather than cooperate [283]. This is exactly what Eq. (27) predicts.

*Discussion and analysis.* The literature suggests three design rules. First, auxiliary tasks should be selected by latent-factor relevance rather than by superficial similarity. Mixed units help because they constrain $\Phi$ and $A$. Dialect or speaker tasks help because they localize $D$ and $S$. Tone and sandhi tasks help because they preserve suprasegmental distinctions that invariance learning may otherwise suppress. Morphology tasks help because they expose reusable word-internal structure. Second, architecture is part of the transfer mechanism. Experts, cross-attention, language IDs, and local attention are ways of reshaping the alignment matrix in Eq. (27) [270-271, 274-275]. Third, schedule matters. Serial task curricula and support-set generation schemes work because they trace smoother paths through task space than naive joint mixing [284-285].

This also clarifies why MTL sometimes outperforms corpus-style pretraining in extremely low-resource conditions and sometimes fails. It wins when one utterance is converted into several aligned constraints, so that Eq. (28) grows faster than the interference terms in Eq. (26). It fails when lexical conflict, writing-system mismatch, or noisy auxiliary targets make the cross-task blocks of Eq. (27) negative. The MUCS challenge is a useful reminder of this limit [286]. Multilingual acoustic sharing can improve likelihood estimation, yet mixed lexica and writing systems can still raise recognition errors if language boundaries remain poorly controlled [287].

*Why this differs from meta-learning.* Meta-learning and MTL solve different residual bottlenecks. Meta-learning optimizes how the model should move after seeing a few examples. Multi-task learning optimizes what shared representation the model should carry before adaptation begins. For tone and morphology, the distinction is practical. If auxiliary tone, F0, or morphological labels are available, MTL usually provides the cleaner signal. If such labels are unavailable but related tasks exist, meta-learning may be the more economical route. The two are compatible and, in current best practice, increasingly complementary.

Although meta-learning and multi-task learning are both subsumed under task transfer, they differ in numerous respects. For clarity, Table 6 is provided at the end of this section to delineate their distinctions.

**Table 6.** Comparison: Meta vs. MTL.

| Dimension | Meta | MTL | Evidence (refs) |
|---|---|---|---|
| Architecture | Agnostic; prompt-tuned | Shared encoder; MoE | [256, 259, 270, 274] |
| Label/unit design | Unified alphabet; priors | Mixed units; LID | [247, 266, 271] |
| Training mechanism | MAML; reweighting | Joint losses; gating | [249, 251, 268-269] |
| Decoding & external LM/post-processing | Beam + LM; prompt bias | LM fusion; consistency | [247, 258, 279, 288] |
| Accent/dialect/genre transfer | Zero-shot accents; invariant reps | Aux LID; genre-aware | [253, 260, 273, 275] |
| Language-specific priors | Tone episodes; inflection paradigms | Tone/F0 heads; morphology heads | [45, 49, 50, 258] |
| Extremely low-resource | Minutes-level; fast adapt | Few-shot SLU; KWS boost | [254, 263, 282, 289] |
| Effect on $C$ and $m$ (upper-bound view) | Shrink $C$; improve adaptation radius | Shrink $C$ + grow $m$ | [250, 267, 290-291] |

### *4.3 Large-Scale Multi-Task Pre-Trained Speech Model*

Large-scale multi-task pre-trained speech models are important for a structural reason. They are among the few model families that can act on the empirical-risk term, the complexity-data trade-off term, and the domain-shift term at the same time. They do not merely push the upstream-hours factor $f_1$ harder. They couple $f_1$, cross-lingual coverage $f_7$, text priors, and task structure into one transferable prior. The historical shift toward larger and broader multilingual, multi-task speech models is summarized in Figure 6. This is why they remain relevant after the eight mainstream techniques begin to saturate.

A convenient way to formalize this point is to treat the pre-training pool as a collection of language-domain-task triplets $\mathcal{T}_{\text{pre}} = \{\tau\}$. Let each $\tau$ contribute a task energy $\mathcal{E}_\tau(\theta)$ and weight $\omega_\tau$. The resulting foundation model induces the aggregate prior

$$P_{\text{fm}}(\theta) \propto P_0(\theta) \exp\left( - \sum_{\tau \in \mathcal{T}_{\text{pre}}} \omega_\tau \mathcal{E}_\tau(\theta) \right) \tag{31}$$

For a low-resource target language $t$, fine-tuning produces a posterior $Q_t$, and the effective complexity term becomes

$$\mathcal{C}_t^{\text{fm}} = \text{KL}(Q_t \,\|\, P_{\text{fm}}) = \text{KL}(Q_t \,\|\, P_0) + \mathbb{E}_{\theta \sim Q_t} \left[ \sum_{\tau \in \mathcal{T}_{\text{pre}}} \omega_\tau \mathcal{E}_\tau(\theta) \right] + \ln Z_{\text{fm}} \tag{32}$$

where $Z_{\text{fm}}$ is the normalizing constant.

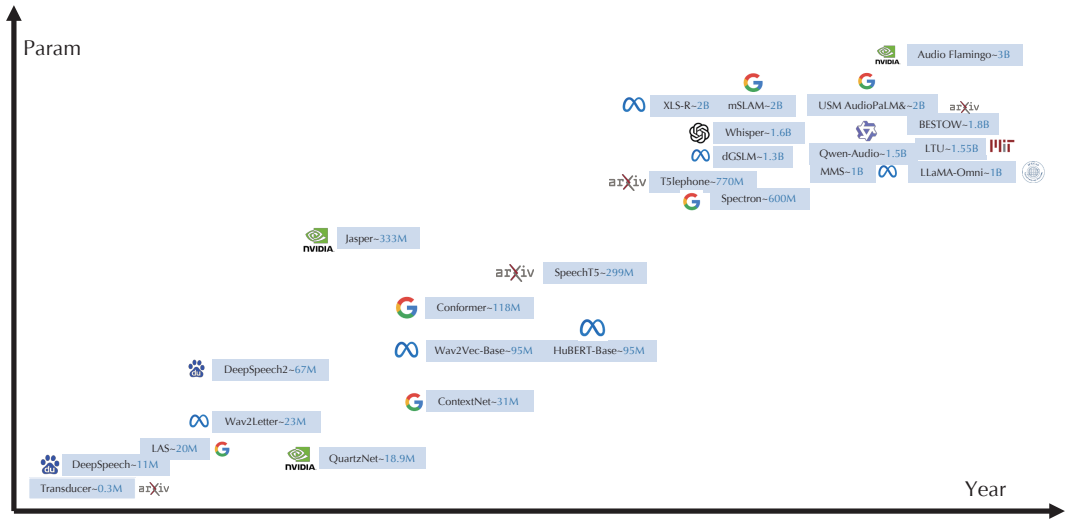



**Figure 6.** Development trend of low-resource ASR models. The key shift is not only toward larger parameter counts, but also toward broader multilingual, multi-task, and multi-domain pre-training that enlarges transferable structure.

The same model family can also turn upstream corpora into *transferable* effective supervision rather than raw hours alone. Let $m_t^{\text{lab}}$ denote labeled target supervision, and let each upstream source $\tau$ contribute raw usable information $m_\tau^{\text{raw}}$. Then a natural target-side approximation is

$$m_t^{\text{tr}} = m_t^{\text{lab}} + \sum_{\tau \in \mathcal{T}_{\text{pre}}} \rho_{\tau \to t} m_\tau^{\text{raw}}, \quad m_\tau^{\text{raw}} = \sum_{\ell} r_{\ell,\tau} H_{\ell,\tau} q_{\ell,\tau} c_{\ell,\tau} \tag{33}$$

where the transfer coefficient $\rho_{\tau \to t} \in [0, 1]$ summarizes how much of the source supervision is reusable for the target. A useful factorization is

$$\rho_{\tau \to t} = \rho_{\text{phn}} \rho_{\text{text}} \rho_{\text{dom}} \rho_{\text{task}} \tag{34}$$

The remaining gain comes from domain coverage. Because the pre-training pool spans many speakers, accents, devices, and recording conditions, the target domain is compared not with one source domain but with a source *mixture*. This yields the bound

$$\Gamma_t^{\text{fm}} \le \sum_k \alpha_k \inf_{\mathbf{w} \in \Delta} \text{Div}_k \left( \sum_{\tau \in \mathcal{T}_{\text{pre}}} w_\tau \mathcal{D}_\tau, \mathcal{D}_t \right) + \Omega(\mathbf{w}) \tag{35}$$

where $\Delta$ is the simplex over pre-training sources and $\Omega(\mathbf{w})$ is the estimation or sparsity cost of the mixture.

Putting these ingredients together gives a foundation-model specialization of the theoretical view:

$$\text{WER}_t \approx R_{\text{noise},t} + A_t \left( \frac{C_t^{\text{fm}}}{m_t^{\text{tr}}} \right)^{1/2} + B_t \Gamma_t^{\text{fm}} \tag{36}$$

where $A_t$ and $B_t$ depend on the target language and evaluation setup.

*Synergy between large-scale speech models and pretraining.* The strongest recent results support the same mechanism from three complementary directions. The first couples speech encoders with large text models or unified speech-text objectives. Representative work includes pretrained language-model fine-tuning, speech-LLM bridges such as Q-Former variants, joint encoder-decoder designs such as BESTOW, text injection for scarce multilingual settings, and unified speech-text pretraining such as SpeechT5, phoneme-aware pretraining, and Transducer-Llama-style vocabulary adaptation [94, 292-299]. Similar conclusions have been reported when wav2vec2 is paired with GPT-4 for few-shot Finnish ASR [300]. These systems matter not simply because the decoder is larger. They matter because they reduce the mismatch between acoustic evidence, pronunciation structure, and text priors.

This reduction can be described by a speech-text bridge energy

$$\mathcal{E}_{\text{bridge}}(\theta) = \lambda_{\text{cm}}\mathbb{E}\left[\| h_\theta^{\text{sp}}(X) - h_\theta^{\text{txt}}(Y) \|_2^2\right] + \lambda_{\text{ph}} H_\theta(\Phi \mid Y,\ h_\theta^{\text{sp}}(X)) + \lambda_{\text{inv}} I_\theta(Z;\ S,\ \Pi,\ U \mid Y) \tag{37}$$

where $h_\theta^{\text{sp}}$ and $h_\theta^{\text{txt}}$ are speech and text representations and $Z$ is the shared latent space. Large-scale joint pre-training lowers this bridge energy by aligning modalities, reducing pronunciation uncertainty, and suppressing nuisance variation.

The second direction scales multilingual speech pre-training itself. Whisper, USM, MMS-style models, and strong adapter-based transfers consistently show that a broad pre-training pool can make four hours, ten hours, or even less than one hour of target labels surprisingly useful [301-307]. Their advantage is not exhausted by saying that $N$ is large. The decisive factor is that $N$ is paired with a large and diverse $\mathcal{T}_{\text{pre}}$, which raises $m_t^{\text{tr}}$ and shrinks $\Gamma_t^{\text{fm}}$ at the same time. This is also why related-language transfer remains important. Moderate models with high $\rho_{\tau \to t}$ can outperform larger but poorly aligned sources [308-310].

The third direction concerns practical routing of the large-model prior into a small adaptation budget. Strong pretrained backbones combined with conventional augmentation remain effective for Arabic, Russian, Portuguese, and other lowresource settings [311-313]. Accent-aware LoRA mixtures, language-family sharing, multilingual tokenizer redesign, and teacher-student compression point to the same principle: the challenge is not only to learn a large prior, but also to expose the right subspace for each target language [314-317]. For tonal or morphologically rich languages, this is precisely where language-specific adapters, sandhi-aware constraints, or morphology-aware output modules become attractive. They act as lightweight corrections to the foundation prior rather than as a rejection of it.

System-level architectural studies reinforce the same message (see Figure 7 for a representative architecture). Multitask large speech models, efficient adaptation modules, and phoneme-augmented reasoning remain effective across spontaneous and highly structured languages [318-323]. The contrast with narrow lightweight systems sharpens this point. Task-specific micro-models can be highly effective when the label space is closed and the domain is tightly controlled, as shown in work on digits, menu ordering, embedded medical ASR, Cantonese edge deployment, aphasia, child speech, and pathological

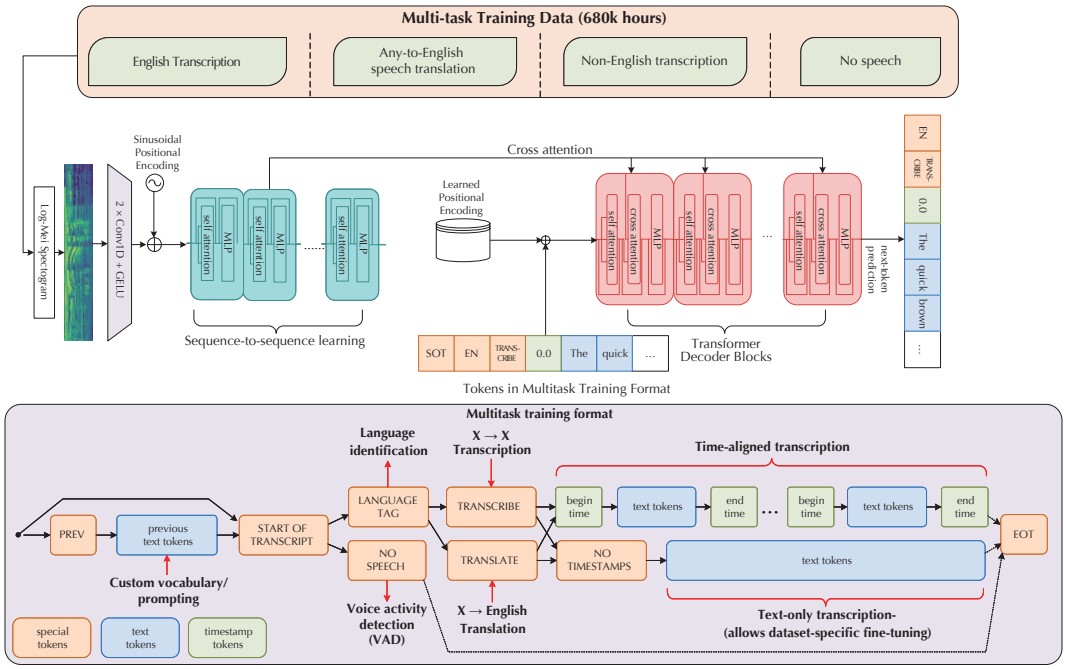



**Figure 7.** A generic architecture of a multilingual multi-task large-scale speech model. The shared acoustic trunk, cross-modal bridge, and task-specific heads together turn multilingual and multi-domain corpora into a reusable target prior.

speech [324-331]. Their efficiency is real, but so is their ceiling: once the target moves outside the narrow domain for which $\Gamma_t$ is artificially small, the absence of a broad transferable prior becomes the dominant bottleneck.

*Large-scale speech model-assisted inference.* The advantage of large-scale speech models also persists at inference time, where strong text priors, retrieval, and cross-modal correction can directly reshape decoding. A useful abstraction is

$$\log \tilde{p}(Y \mid X,\ c) = \log p_{\text{asr}}(Y \mid X) + \lambda \log p_{\text{llm}}(Y \mid c) + \gamma\ \mathrm{s}_{\text{ret}}(Y,\ c) \tag{38}$$

where $c$ may denote prompts, retrieved documents, terminology lists, dialogue state, or other side information. This equation explains why LLM-coupled decoding, rescoring, and retrieval-assisted correction often help most in low-resource settings: they sharpen the text prior where the first-pass acoustic posterior is still diffuse.

Recent systems illustrate this mechanism from several angles. Joint inference pipelines that couple Whisper-like encoders with one or more LLMs, residual state injection methods such as SALSA, and staged speech-to-LLM systems all show that language-side capacity can compensate for sparse acoustic supervision [332-335]. Streaming and real-time variants reach a similar conclusion via online distillation

and streaming adaptations of Whisper [336-337]. Post-correction systems make the argument even clearer because they operate after the acoustic model is fixed [338-342]. If $\pi_{\text{fix}}$ is the probability that a true error is corrected and $\pi_{\text{fp}}$ is the probability that a correct token is damaged, then the residual risk obeys

$$R_{\text{post}} \leq (1-\pi_{\text{fix}})R_{\text{lst}} + \pi_{\text{fp}}(1-R_{\text{lst}}) \tag{39}$$

Cross-lingual and dialectal evaluations support the same reading. Whisper-family and related large models have been adapted to Kurdish, Tamil dialects, Arabic dialects, Turkish, Amharic, Urdu, and other low-resource settings with broadly positive results [343-349]. Yet related-language corpora, script projection, and external language models still matter strongly in Livvi-Karelian, Mongolian, Indo-Aryan, and Dravidian settings [308-310, 350]. Similar lessons also appear in legal terminology detection and competition SLU pipelines, where an ASR + NLU/LLM decomposition remains preferable when direct end-to-end learning of domain semantics is too expensive [351-352].

*Remarks.* The main lesson is not that larger models dominate unconditionally. The stronger and more useful claim is narrower: large-scale multi-task pre-trained speech models are theoretically attractive because they are among the few approaches that can reduce $\mathcal{C}_{\text{eff}}$, increase transferable $m_{\text{eff}}$, and shrink $\Gamma$ in one pipeline. This advantage also explains their sensitivity to data quality. As model capacity grows, poor synthetic data, noisy pseudo-labels, and weak teachers become more dangerous because the model has enough capacity to fit them well. A simple quality-adjusted approximation is

$$m_{\text{pre}}^{\text{usable}}(N) = \sum_{\tau \in \mathcal{T}_{\text{pre}}} \rho_{\tau \to t} H_\tau (2q_\tau - 1)_+^{\nu(N)}, \quad \nu'(N) > 0 \tag{40}$$

which says that low-quality sources are discounted more sharply as model size $N$ grows [353-354].

Deployment introduces a second constraint. Hallucination, forgetting, latency, and memory cost all become part of the real objective [299, 354]. Compression and efficient adaptation are therefore better seen as constrained risk minimization:

$$\min_{\mathcal{A}} \left[ \text{WER}_t(\mathcal{A}) + \lambda_{\text{lat}} \text{RTF}(\mathcal{A}) + \lambda_{\text{mem}} \text{Mem}(\mathcal{A}) + \lambda_{\text{hall}} \text{Hall}(\mathcal{A}) \right] \tag{41}$$

where $\mathcal{A}$ may include LoRA, SVD compression, quantization, expert mixing, task-vector merging, or selective attention merging [355-359]. The practical pipeline is therefore not simply "bigger is better." It is better described as "scale → filter → compress → adapt."

Taken together, the literature supports a sharper conclusion than the generic claim that foundation models help lowresource ASR. Their value lies in the way they reorganize transfer. Multilingual sharing enlarges coverage. Multi-task supervision turns that coverage into structured constraints. Cross-modal coupling sharpens the bridge between acoustics and text. Efficient adaptation preserves most of this prior under a small update budget. This is precisely why large-scale multi-task pre-trained speech models remain one of the most plausible routes beyond the current saturation point in low-resource ASR.

### 4.4 Corpus Augmentation Techniques

#### 4.4.1 Data Synthesis and Augmentation

As discussed above, the scale and quality of audio corpora remain primary determinants of model performance. Corpus augmentation is therefore a research direction that is both orthogonal and complementary to transfer-learning design and model scaling. A concise summary of the main augmentation routes is presented in Table 7.

**Table 7.** Approaches for Expanding Audio Training Corpora.

| Approach | Core idea | Primary techniques |
|---|---|---|
| Synthetic corpus generation | Synthetic speech via deep models | TTS; GAN-based TTS |
| Multimodal transfer | Exploit cross-modal correlations | Cross-modal masked self-supervised |
| Corpus development | Human-led corpus expansion | Crowdsourcing; automatic annotation |

Current augmentation methods span a full spectrum, from better use of existing resources to outright generation of new data. On one end of this spectrum, self-supervised learning has enabled domain-aware augmentation strategies. Automatic selection of augmentation policies via Hilbert-Schmidt dependence measures can reduce the impact of domain mismatch [360]. In cross-lingual settings, multi-encoder single-decoder architectures can map source-language audio into target-language "ciphertext" and thereby create synthetic transfer paths for data-scarce languages [361]. Other work builds universal speech units from self-supervised representations and then performs efficient cross-lingual synthesis through unit concatenation, reducing dependence on conventional full TTS pipelines [362-363].

TTS-based augmentation forms the most direct bridge between text and speech. A representative workflow of TTS-based data augmentation is shown in Figure 8. Different studies target different failure modes. Some address the distribution gap between synthetic and real speech through discrete vq-wav2vec features [364]. Others synthesize pathological or dysarthric speech from limited patient recordings [365].

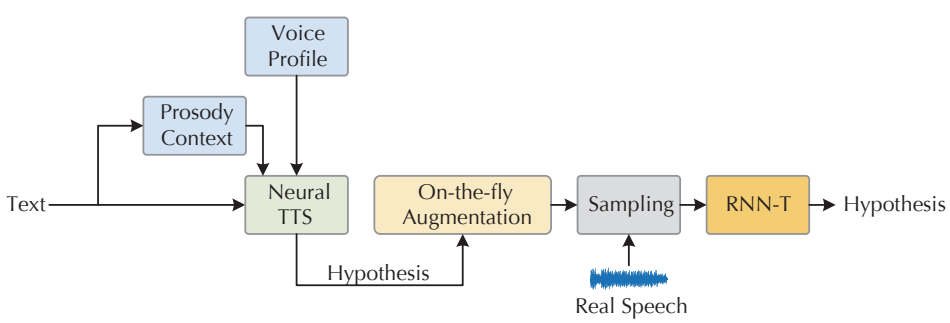



**Figure 8.** Workflow of TTS-based Data Augmentation.

Multilingual TTS has also been used to generate multi-style dialectal speech and to simulate accented speech in order to reduce recognition bias [366-367]. In extremely low-resource settings, text augmentation can be coupled with TTS to expand both lexical and acoustic diversity [368-369].

Adversarial learning extends this idea by better capturing subtle speech characteristics. A generic GAN-based TTS data-augmentation framework is illustrated in Figure 9. In pathological speech, VAEGAN systems disentangle content and speaker factors to synthesize higher-quality disordered speech [370]. In stuttering speech, adversarial and signal-level methods can improve recognition by modeling the fine structure of dysfluency [371]. Similar GAN-based approaches have been used to bridge adult-child mismatch in both monolingual and mixed recognition scenarios [372-373]. Search-oriented systems further show that contrastive text selection and fine-grained GAN synthesis can deliver practical gains in real applications [374].

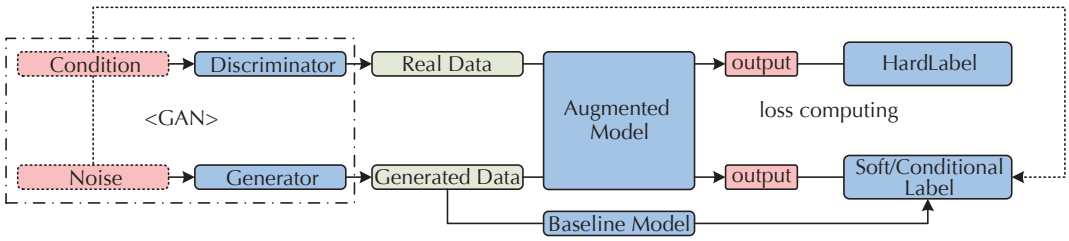



**Figure 9.** GAN-based TTS Data Augmentation.

### *4.4.2 Cross-Modal Information Integration*

Cross-modal supplementation has also advanced rapidly. The central idea is not merely to "borrow more data" from another modality. It is to translate visual, textual, articulatory, or large-model information into constraints that can replace or reinforce scarce labeled speech. A complete multilingual pretraining and transfer pipeline of this kind has already brought an Italian system with only 160 hours of training data close to production-level accuracy [375]. Similar multimodal strategies based on speech and text seed data with cross-attention have improved recognition in extremely low-resource dialects such as Latin Cuengh [376].

*Information augmentation.* Several methods weaken dependence on transcription by changing the supervision channel. ASR-free spoken retrieval aligns speech and text in a dual-encoder space and avoids the propagation of high WER, although single-vector retrieval introduces a capacity bottleneck for long queries [377]. Subtitle time alignment injects audio self-supervision and cross-domain transfer into temporal localization, which helps even when subtitle and audio languages differ [378]. Shared acoustic-semantic representations can turn large volumes of untranscribed audio into supervision for topic identification [379]. Speech-to-image conversion for CNN-based keyword spotting explores a different visual encoding of acoustic variation, but also reveals sensitivity to feature choice in transfer settings [380]. Audio-visual HMM adaptation and masked cross-modal self-supervision both support the same general

lesson: if another modality approximates the task-conditional structure closely enough, it can substitute for part of the annotation burden [381-382].

*Cross-modal substitution.* A second route treats visual or articulatory modalities as explicit acoustic cues. This is particularly useful in noisy or endangered-language settings, where visible articulators can compensate for missing or degraded acoustic evidence. Work on Cantonese in-car speech, Tujia sequence modeling, interleaved multimodal finetuning, and EMA-based cross-lingual articulatory transfer shows a clear progression: visual reinforcement improves noise robustness, sequence models recover temporal dependencies, and careful training schedules prevent catastrophic forgetting of cross-modal capability [147, 383-385].

*Coupling with multi-task foundation models.* Large multilingual and multi-task models make cross-modal transfer more powerful because they distill higher-level semantic, translation, and dialogue knowledge into speech processing. AVHuBERT pre-trained in English can support Persian AVSR and viseme discovery, illustrating both cross-lingual reuse and pipeline automation [386]. Parameter-efficient adaptation that first improves machine translation capacity with text and only then incorporates limited ASR data is particularly attractive for extremely low-resource languages [387]. Broader evaluations also show that the transfer boundary of multilingual self-supervised encoders depends jointly on language relatedness, task semantics, and corpus scale [388-391].

On the text side, supplementation and weak-label propagation follow the same logic. Book-dialogue embeddings can be transferred to low-resource emotion recognition [392]. Weak labels generated with English RoBERTa can be projected to multilingual audio through aligned text resources such as the Bible [393]. New multimodal emotion datasets in Moroccan Arabic and Sindhi further show both the promise and the boundary of lexical-prosodic transfer [394-395]. Engineering challenge settings, such as staged Cantonese in-car noise benchmarks, are useful because they expose the real trade-off between controllable and variable noise conditions and force a realistic evaluation of pretraining, augmentation, and modality fusion [396].

#### *4.4.3 Advancing Corpus Development*

Low-resource corpus development is advancing rapidly in data collection, annotation quality control, and multilingual transfer. Current work can be grouped into three broad routes.

The first route expands usable data through multi-source text and large-scale pretraining. Telugu ASR is a strong example: large-scale collection from Wikipedia and local websites created a text corpus of more than eight million sentences, which then supported deep-model pretraining and downstream recognition [397]. CoVoST 2 built a translation corpus spanning 21 languages and 2,880 hours of speech through crowdsourced translation and language-model scoring [398]. FLEURS provided few-shot evaluation across 102 languages and highlighted the importance of script similarity and pretraining coverage [399]. mTEDx extended this trend toward scalable multilingual ASR and speech translation [400].

The second route builds specialized corpora for specific dialects, speech communities, or application conditions. Swahili numerical speech was recorded under noisy conditions and included both native and non-native speakers in order to reflect actual use scenarios [401]. The Lhasa Tibetan NICT-Tib1 corpus contributed 33.5 hours of high-quality recordings and demonstrated its usefulness through multilingual model training [402]. A Congo-language corpus combined read and broadcast speech across four languages, including Lingala, and laid the foundation for multilingual acoustic modeling in that region [403].

The third route improves corpus value through language models and phoneme alignment. Uzbek ASR has benefited from very large crawled text corpora paired with moderate-sized speech collections, which substantially reduced test-set WER after LM integration [404]. Multilingual phoneme datasets built through automatic alignment and real-time correction provide a new path for languages without stable orthography or with underdeveloped writing standards [405].

Taken together, these strands show that corpus construction is no longer just a matter of collecting more hours. Multisource mining and pretraining support cross-language transfer; specialized corpora address dialectal and domain-specific gaps; and alignment or LM techniques increase the value of each collected utterance. Future work will likely continue along three fronts: script-aware pretraining, robust collection under real noise, and broader open collaboration, especially for far-field speech, children's speech, and other underrepresented conditions.

### *4.5 Synthesis Roadmap and Open Questions*

This section concludes the paper by drawing together its main arguments and distilling, from the full body of evidence reviewed above, a forward-looking research blueprint for the next stage of low-resource ASR. The central conclusion is that the field is approaching a transition from *stacking transfer techniques* to *orchestrating transfer*. The core question is no longer only whether knowledge can be borrowed from other languages, tasks, or models, but whether that borrowed knowledge can be allocated to the *right latent factors* under an extremely small target-language budget.

Throughout the paper, the recurrent failure mode was not the absence of transfer, but the *misallocation* of transfer. Generic invariance can erase lexical tone; frequency-driven subwording can blur morphology; full fine-tuning can spend scarce updates on the wrong layers; and self-training can amplify exactly the low-frequency errors that matter most in deployment. Read through Eq. (13), the real scarce resource is therefore not raw parameters alone, but *reliable targetspecific plasticity*: the small set of trainable degrees of freedom that should reduce $C_{\text{eff}}$, enlarge $m_{\text{eff}}$, and control $\Gamma$ without destroying useful prior structure.

*From three frontiers to one convergent stack.* Under this view, the three directions that appear most strategic in current research—meta-learning, multi-task learning, and large-scale multi-task speech models—should not be treated as parallel strategies. They are better understood as three layers of the same emerging system.

- **Large-scale multi-task speech models** are increasingly the *prior substrate*. Their main role is to absorb multilingual, multi-domain, and partially multimodal regularities before the target language is seen. In the next stage, their value will lie less in raw parameter count than in how well they expose reusable internal structure to later adaptation.
- **Multi-task learning** is becoming the *structure extractor*. Its future role is not merely to add more auxiliary losses, but to factorize the latent causes of recognition difficulty—segmental content, tone and prosody, morphology, speaker and dialect, discourse, and semantics—so that each utterance yields several aligned constraints rather than one transcript only.
- **Meta-learning** is becoming the *adaptation controller*. Rather than serving only as a better initialization, it is likely to decide where adaptation should occur, which auxiliary tasks should be activated, which unlabeled samples should be trusted, and how a tiny trainable budget should be distributed across layers, experts, or adapters.

The important prediction is therefore a new division of labor. Foundation-scale models will supply a rich but overcomplete prior; multi-task factorization will specify which structure must remain explicit; and meta-learning will regulate how that prior moves under severe label scarcity. In such a design, PEFT becomes the actuation interface, self-training becomes the target-evidence amplifier, retrieval and language-model coupling become text-side correctors, and multimodal transfer becomes a substitute signal when speech alone is insufficient. These techniques, which remain insufficiently integrated and relatively isolated in current research, are likely to become modules in a single closed-loop transfer system.

This also changes the status of language-specific priors. Tone-aware constraints, sandhi modeling, morphology-aware decoders, and dialect-aware modules are not side patches added after generic transfer. They are criteria for deciding which latent variables must remain explicit and which parts of the model deserve plasticity. The boundary between corpus-derived, model-encoded, and task-structural transfer thus becomes an implementation boundary rather than a strategic one.

*A roadmap beyond linear pretrain-adapt pipelines.* Figure 10 summarizes the resulting roadmap. The figure should be read from left to right, but also as an outer feedback loop. Its message is that the next generation of low-resource ASR is unlikely to be a longer pipeline of loosely coupled tricks. It will more likely be a layered system in which prior construction, structure exposure, controlled adaptation, and trustworthy evidence expansion are co-designed.

1. **Foundation prior construction.** The entry point remains large-scale multilingual and multi-task pretraining, potentially augmented by text, translation, and multimodal corpora. The objective at this stage is breadth: maximize transferable coverage before the target language is encountered.
2. **Latent-factor exposure.** Before adaptation, the model should be equipped with a structured task bank that makes the under-modeled factors of low-resource ASR explicit. For some languages, the key factors are tone categories, tone sandhi, or F0 dynamics; for others, they are morpheme boundaries, inflectional bundles, code-switch transitions, speaker and dialect cues, or domain-specific lexical constraints. The point is not to hand-engineer every language, but to prevent the most consequential low-frequency structure from remaining invisible.

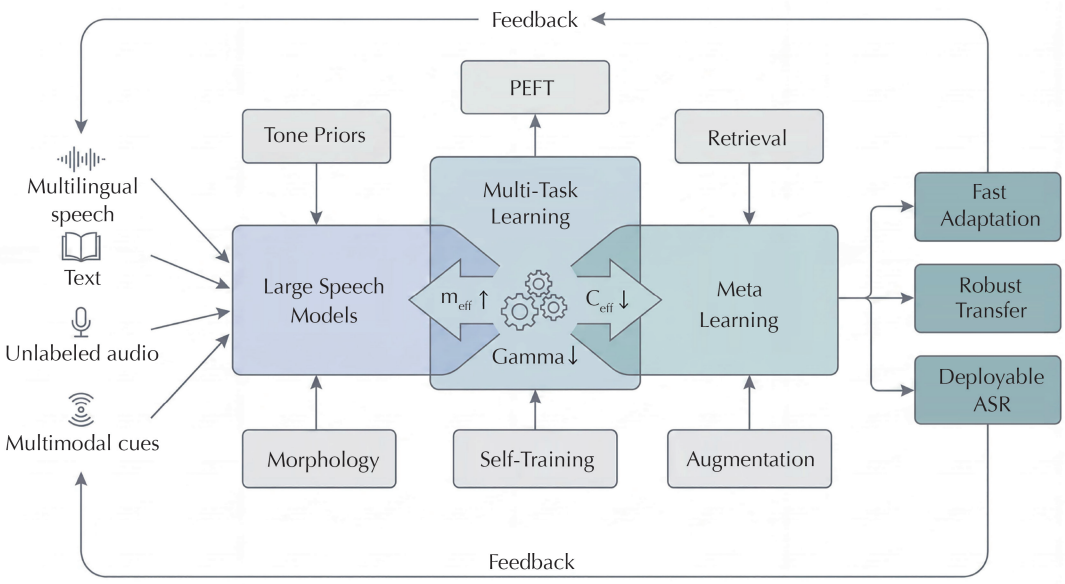



**Figure 10.** A forward blueprint for low-resource ASR beyond isolated techniques. Foundation-scale speech models supply a multilingual and multi-domain prior; multi-task factorization exposes the latent structure that generic transfer tends to under-model; meta-controlled adaptation allocates a tiny update budget to the right layers, experts, and auxiliary tasks; and a trust-aware feedback loop couples self-training, retrieval, post-correction, and deployment. The goal of the whole stack is joint: enlarge $m_{eff}$, reduce $C_{eff}$, and control $\Gamma$ simultaneously rather than one term at a time.

3. **Dynamic low-budget specialization.** Adaptation should then be treated as a controlled allocation problem. Instead of updating a model uniformly, the system should learn where to spend its limited plasticity: in routing, in adapters, in low-rank updates, in task weights, or in a small number of decoder-side corrections. This is the natural meeting point of meta-learning and PEFT.
4. **Trustworthy evidence expansion and inference.** After the first adaptation pass, the system should enlarge usable supervision through self-training, retrieval-assisted decoding, post-correction, and targeted augmentation. Crucially, this stage must be trust-aware. Otherwise, unlabeled audio increases nominal data volume while worsening long-tail errors.
5. **Deployment feedback.** The final stage is not deployment as an endpoint, but deployment as a data-collection and diagnosis loop. Real failures expose which latent factors were not sufficiently modeled, which confidence signals were miscalibrated, and which parts of the adaptation budget were misused. Those signals should feed back into task design, adapter placement, and corpus construction.

This roadmap implies a methodological shift. Future systems should be evaluated not only by average WER, but also by *where* their remaining errors concentrate: tone minimal pairs, sandhi contexts, rare inflected forms, dialect boundaries, code-switch points, child or elderly speech, and retrieval-sensitive terminology. Once foundation models become strong enough, mean performance alone becomes too coarse to guide the next round of design.

*A sharper frontier hypothesis*. The next major gain in low-resource ASR is unlikely to come from scale alone, and equally unlikely to come from language-specific structure in isolation. It is more likely to come from *selective plasticity*: a regime in which (i) large models provide a broad prior, (ii) multi-task objectives keep the right latent variables explicit, and (iii) meta-controlled adaptation decides which small subset of the model is allowed to move for each language, domain, and data regime. In other words, the frontier is shifting from *pretraining strength* to *transfer governance*.

This also explains why the three frontier directions emphasized in this paper will increasingly co-evolve, rather than remain isolated or be framed as if they must be traded off against one another. A stronger foundation model makes task factorization more valuable, because the model already contains many reusable features but does not know which ones should remain active. Better multi-task factorization makes meta-learning more effective, because adaptation episodes become aligned with genuine linguistic structure rather than arbitrary label sets. Better meta-learning, in turn, makes foundation models safer to deploy in one-hour or few-hour regimes, because the model no longer needs to relearn an entire language from scratch; it only needs to discover how to expose and specialize the prior it already contains.

*Three precise and answerable research questions*. The following questions are intended to be narrow enough for controlled study, yet deep enough to shape the next stage of the field.

1. **Where should adaptation live in a foundation-scale model?** For a fixed trainable budget and a fixed target language with extremely limited labels, is it better to meta-learn the initialization, the routing policy, the adapter placement, or the auxiliary-task weights? This question is especially important for tonal and morphologically rich languages, where the optimal location of plasticity may differ sharply from that of non-tonal or morphologically simpler languages. A convincing answer should report not only WER, but also tone-confusion rates, sandhi consistency, and rare-form generalization.
2. **Which explicit factorization yields the largest information gain per utterance?** Given the same backbone and the same annotation budget, which structured task bank contributes the greatest increase in usable supervision: phonetic plus tone tasks, phonetic plus morphology tasks, speaker/dialect-aware tasks, semantic tasks, or a dynamically selected combination of them? The core issue is not auxiliary-task quantity, but whether one additional minute of annotation produces a larger reduction in $C_{\text{eff}}$ and a larger increase in $m_{\text{eff}}$ than generic transcript-only training.
3. **Can the training value of an unlabeled utterance be predicted before it is used?** Rather than filtering pseudolabels by confidence alone, can one estimate whether an utterance will improve or harm the target model by combining acoustic confidence, retrieval agreement, and tone- or morphology-aware consistency checks? If this is possible, self-training becomes a selective expansion mechanism for $m_{\text{eff}}$ rather than a blind increase in nominal data volume. This is the decisive test for whether unlabeled speech can be turned into trustworthy structure instead of amplified bias.

Taken together, these questions define a more unified agenda than "bigger models" versus "better inductive bias." The most promising systems will likely be those that learn a broad prior at foundation scale, expose linguistically meaningful structure through multi-task objectives, and then adapt that prior

through meta-controlled, budget-aware, and trust-aware specialization. That is the sense in which the future of low-resource ASR appears convergent: not toward one dominant trick, but toward a coordinated architecture in which scale, structure, and adaptation cease to be separate choices.

## AUTHOR CONTRIBUTIONS

**Donghong Qin** was responsible for funding acquisition, project administration, resources, supervision, and writingreview. **Zhuo Chen** was responsible for conceptualization, data curation, formal analysis, investigation, validation, visualization, and writing-original draft preparation. **Xiao Liang** was responsible for correspondence, supervision, validation, and writing-review and editing. **Fengbo Bai** was responsible for supervision and writing-review. All authors have read and agreed to the published version of the manuscript.

## ACKNOWLEDGEMENTS

This paper was funded by the Strategic Research and Consulting Project of the Chinese Academy of Engineering (2025DFZD5202), the Central Government-Guided Local Science and Technology Development Fund of Guangxi Zhuang Autonomous Region (Grant No. Gui Ke ZY24212045), and the Guangxi Key Research and Development Program (Grant No. Gui Ke AB25069456).

The authors sincerely thank the reviewers for their constructive and insightful comments, which helped to substantially improve the clarity, rigor, and overall quality of this manuscript. The authors are also grateful to the Editor-in-Chief, the Guest Editors, and the editorial office staff for their timely, patient, and professional communication throughout the revision and publication process.

## REFERENCES

[1] S. Singh, F. Hou, and R. Wang, "A novel self-training approach for low-resource speech recognition," in *Interspeech 2023*, 2023, pp. 1588–1592.

[2] N. H. Gebreegziabher and A. Nürnberger, "An amharic syllable-based speech corpus for continuous speech recognition," in *Statistical Language and Speech Processing*, C. Martín-Vide, M. Purver, and S. Pollak, Eds. Cham: Springer International Publishing, 2019, pp. 177–187.

[3] K. Raju, A. V, R. Lish, and J. Mathew, "Snow mountain: Dataset of audio recordings of the bible in low resource languages," *CoRR*, vol. abs/2206.01205, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.01205.

[4] A. Imran and K. Sunil, "Speech recognition for resource deficient languages using frugal speech corpus," in *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*, 2012, pp. 750–755.

[5] I. Kipyatkova and I. Kagirov, "Analytical review of methods for solving data scarcity issues regarding elaboration of automatic speech recognition systems for low-resource languages," *Informatics and Automation*, vol. 21, no. 4, pp. 678–709, 2022. [Online]. Available: http://mi.mathnet.ru/trspy1205.

[6] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639313000988.

[7] K. Yufune, T. Koriyama, S. Takamichi, and H. Saruwatari, "Accent modeling of low-resourced dialect in pitch accent language using variational autoencoder," in *11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 189–194.

[8] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Columbia University, USA, 2011, aAI3450188.

[9] B. T. Meyer, T. Jrgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition depending on speech-intrinsic variabilitya)," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3126–3141, 11 2010. [Online]. Available: https://doi.org/10.1121/1.3493450.

[10] N. F. Chen, B. P. Lim, V. H. Do, V. T. Pham, C. Ni, H. Xu, M. HasegawaJohnson, W. Chen, X. Xiao, S. Sivadas, E. S. Chng, B. Ma, and H. Li, "Low-resource spoken keyword search strategies in georgian inspired by distinctive feature theory," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1322–1327.

[11] A. Gupta, "On building spoken language understanding systems for low resourced languages," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, G. Nicolai and E. Chodroff, Eds. Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 1–11. [Online]. Available: https://aclanthology.org/2022.sigmorphon-1.1/.

[12] J. H. McWhorter, "The worlds simplest grammars are creole grammars," *Linguistic Typology*, vol. 5, no. 2–3, pp. 125–166, 2001. [Online]. Available: https://doi.org/10.1515/lity.2001.001.

[13] R. Eskander, J. Klavans, and S. Muresan, "Unsupervised morphological segmentation for low-resource polysynthetic languages," in *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, G. Nicolai and R. Cotterell, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 189–195. [Online]. Available: https://aclanthology.org/W19-4222/.

[14] K. Peterson, A. Tong, and Y. Yu, "Openasr20: An open challenge for automatic speech recognition of conversational telephone speech in low-resource languages," in *Interspeech 2021*, 2021, pp. 4324–4328.

[15] T. Reitmaier, E. Wallington, D. Kalarikalayil Raju, O. Klejch, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Opportunities and challenges of automatic speech recognition systems for low-resource language speakers," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1–17. [Online]. Available: https://doi.org/10.1145/3491102.3517639.

[16] K. Fatehi, M. Torres Torres, and A. Kucukyilmaz, "An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments," *Speech Communication*, vol. 167, p. 103151, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639324001225.

[17] Å. Lepak, K. Radzikowski, R. Nowak, and K. J. Piczak, "Generalisation gap of keyword spotters in a cross-speaker low-resource scenario," *Sensors*, vol. 21, no. 24, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/24/8313.

[18] M. Karfidt, M. K. Baskar, K. Vesel, F. Grzl, L. Burget, and J. ernock, "Analysis of multilingual blstm acoustic model on low and high resource languages," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5789–5793.

[19] Y. Zhang, E. Chuangsuwanich, J. Glass, and D. Yu, "Prediction-adaptation-correction recurrent neural networks for low-resource language speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5415–5419.

[20] V. M. Shetty and M. Sagaya Mary N.J., "Improving the performance of transformer based low resource speech recognition for indian languages," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8279–8283.

[21] S. Feng, P. Åelasko, L. Moro-Velzquez, A. Abavisani, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "How phonotactics affect multilingual and zero-shot asr performance," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7238–7242.

[22] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estve, and L. Besacier, "Investigating self-supervised pre-training for end-to-end speech translation," in *Interspeech 2020*, 2020, pp. 1466–1470.

[23] H. Xue, Q. Shao, P. Chen, P. Guo, L. Xie, and J. Liu, "Tranusr: Phoneme-to-word transcoder based unified speech representation learning for cross-lingual speech recognition," in *Interspeech 2023*, 2023, pp. 216–220.

[24] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Transfer ability of monolingual wav2vec2.0 for low-resource speech recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–6.

[25] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 914–921.

[26] A. Ullah, A. Ragano, and A. Hines, "Improving phoneme recognition with augmented autoregressive predictive coding," in *2023 34th Irish Signals and Systems Conference (ISSC)*, 2023, pp. 1–6.

[27] L.-M. Lam-Yee-Mui, L. O. Yang, and O. Klejch, "Comparing self-supervised pre-training and semi-supervised training for speech recognition in languages with weak language models," in *Interspeech 2023*, 2023, pp. 87–91.

[28] L. Sun, N. Yolwas, and L. Jiang, "A method improves speech recognition with contrastive learning in low-resource languages," *Applied Sciences*, vol. 13, no. 8, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/8/4836.

[29] Z. Zhang, W. Wang, and Y. Qian, "Fast and efficient multilingual self-supervised pre-training for low-resource speech recognition," in *Interspeech 2023*, 2023, pp. 2248–2252.

[30] H. Zhu, L. Wang, G. Cheng, J. Wang, P. Zhang, and Y. Yan, "Wav2vec-s: Semi-supervised pre-training for low-resource asr," in *Interspeech 2022*, 2022, pp. 4870–4874.

[31] A. Wu, C. Wang, J. Pino, and J. Gu, "Self-supervised representations improve end-to-end speech translation," in *Interspeech 2020*, 2020, pp. 1491–1495.

[32] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023, pp. 1416–1429.

[33] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?" in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6533–6537.

[34] 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 *Interspeech 2023*, 2023, pp. 1384–1388.

[35] P. Vieting, C. Lscher, J. Dierkes, R. Schlter, and H. Ney, "Efficient utilization of large pre-trained models for low resource asr," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023, pp. 1–5.

[36] K. Chang, Y.-H. Chou, J. Shi, H.-M. Chen, N. Holliday, O. Scharenborg, and D. R. Mortensen, "Self-supervised speech representations still struggle with african american vernacular english," in *Interspeech 2024*, 2024, pp. 4643–4647.

[37] J. Zhao and W.-Q. Zhang, "Improving automatic speech recognition performance for low-resource languages with self-supervised models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1227–1241, 2022.

[38] R. Fan, "Improving the accuracy and inference efficiency for Low-Resource automatic speech recognition," Ph.D. dissertation, University of California, Los Angeles, United States-California, 2024.

[39] K. Fatehi and A. Kucukyilmaz, "Labert: A combination of local aggregation and self-supervised speech representation learning for detecting informative hidden units in low-resource asr systems," in *Interspeech 2023*, 2023, pp. 211–215.

[40] T. Srivastava, J. Shi, W. Chen, and S. Watanabe, "Effuse: Efficient self-supervised feature fusion for e2e asr in low resource and multilingual scenarios," in *Interspeech 2024*, 2024, pp. 3989–3993.

[41] 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," in *Interspeech 2020*, 2020, pp. 4751–4755.

[42] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[43] Y. Karunanayake, U. Thayasivam, and S. Ranathunga, "Transfer learning based free-form speech command classification for low-resource languages," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, F. Alva-Manchego, E. Choi, and D. Khashabi, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 288–294. [Online]. Available: https://aclanthology.org/P19-2040/.

[44] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," in *Interspeech 2022*, 2022, pp. 2113–2117.

[45] H. Mohiuddin, Z. Ahmed, M. Kasi, and B. Kasi, "Urduspeakxlsr: Multilingual model for urdu speech recognition," in *2023 18th International Conference on Emerging Technologies (ICET)*, 2023, pp. 217–221.

[46] F. Wang, Y. Yang, and J. Yang, "End-to-end myanmar speech recognition with human-machine cooperation," in *2022 International Conference on Asian Language Processing (IALP)*, 2022, pp. 156–161.

[47] M. Y. Tachbelie, S. T. Abate, and T. Schultz, "Development of multilingual asr using globalphone for less-resourced languages: The case of ethiopian languages," in *Interspeech 2020*, 2020, pp. 1032–1036.

[48] S. Bhosale, I. Sheikh, S. H. Dumpala, and S. K. Kopparapu, "End-to-end spoken language understanding: Bootstrapping in low resource scenarios," in *Interspeech 2019*, 2019, pp. 1188–1192.

[49] N. Fathima, T. Patel, M. C, and A. Iyengar, "Tdnn-based multilingual speech recognition system for low resource indian languages," in *Interspeech 2018*, 2018, pp. 3197–3201.

[50] M. A, B. Pilar, and R. Angarai Ganesan, "Subword dictionary learning and segmentation for expanding the vocabulary of automatic speech recognition in tamil and kannada," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 24, no. 1, Jan. 2025. [Online]. Available: https://doi.org/10.1145/3705312.

[51] R. Sahraeian and D. Van Compernolle, "Using weighted model averaging in distributed multilingual dnns to improve low resource asr," *Procedia Computer Science*, vol. 81, pp. 152–158, 2016, sLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09–12 May 2016 Yogyakarta, Indonesia. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916300576.

[52] S. Qin, L. Wang, S. Li, J. Dang, and L. Pan, "Improving low-resource tibetan end-to-end ASR by multilingual and multilevel unit modeling," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 2, Jan. 2022.

[53] V. Gupta and G. Boulianne, "Progress in multilingual speech recognition for low resource languages Kurmanji Kurdish, Cree and inuktut," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6420–6428. [Online]. Available: https://aclanthology.org/2022.lrec-1.689/.

[54] A. Bekarystankyzy, A. Razaque, and O. Mamyrbayev, "Integrated end-to-end multilingual method for low-resource agglutinative languages using cyrillic scripts," *Journal of Industrial Information Integration*, vol. 43, p. 100750, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452414X24001936.

[55] J. Luo, J. Wang, N. Cheng, Z. Zheng, and J. Xiao, "Adaptive activation network for low resource multilingual speech recognition," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–7.

[56] Y. Lu, M. Huang, X. Qu, P. Wei, and Z. Ma, "Language adaptive cross-lingual speech representation learning with sparse sharing sub-networks," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6882–6886.

[57] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual end-to-end speech translation," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 570–577.

[58] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[59] A. S. Kirkedal and Y.-J. Kim, "Multilingual deep neural network training using cyclical learning rate," in *Interspeech 2018*, 2018, pp. 2933–2937.

[60] S. Thomas, K. Audhkhasi, and B. Kingsbury, "Transliteration based data augmentation for training multilingual asr acoustic models in low resource settings," in *Interspeech 2020*, 2020, pp. 4736–4740.

[61] R. Sahraeian and D. Van Compernolle, "Cross-entropy training of dnn ensemble acoustic models for low-resource asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1991–2001, 2018.

[62] C.-I. J. Lai, Y. Zhang, A. H. Liu, S. Chang, Y.-L. Liao, Y.-S. Chuang, K. Qian, S. Khurana, D. Cox, and J. Glass, "Parp: Prune, adjust and re-prune for self-supervised speech recognition," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 21256–21272. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/b17c0907e67d868b4e0feb43dbbe6f11-Paper.pdf.

[63] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for lvcsr," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4955–4959.

[64] G. Deekshitha and L. Mary, "Multilingual spoken term detection: a review," *International Journal of Speech Technology*, vol. 23, no. 3, pp. 653–667, Sep. 2020.

[65] R. Sahraeian and D. Van Compernolle, "Crosslingual and multilingual speech recognition based on the speech manifold," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2301–2312, 2017.

[66] X. Yue, G. Lee, E. Ylmaz, F. Deng, and H. Li, "End-to-end code-switching asr for low-resourced language pairs," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 972–979.

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

[68] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021, pp. 27826–27839.

[69] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "slimipl: Language-model-free iterative pseudo-labeling," in *Interspeech 2021*, 2021, pp. 741–745.

[70] A. Przezdziak, "Optimizing speech recognition for low-resource languages: Northern sotho," Master's thesis, Uppsala University, Department of Linguistics and Philology, 2024.

[71] H. Xu, H. Su, C. Ni, X. Xiao, H. Huang, E. S. Chng, and H. Li, "Semi-supervised and cross-lingual knowledge transfer learnings for dnn hybrid acoustic models under low-resource conditions," in *Interspeech 2016*, 2016, pp. 1315–1319.

[72] 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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80/.

[73] S. Khurana, A. Laurent, W.-N. Hsu, J. Chorowski, A. Lancucki, R. Marxer, and J. Glass, "A convolutional deep markov model for unsupervised speech representation learning," in *Interspeech 2020*, 2020, pp. 3790–3794.

[74] W. Meng and N. Yolwas, "A study of speech recognition for kazakh based on unsupervised pre-training," *Sensors*, vol. 23, no. 2, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/2/870.

[75] C. Y. Kwok, J. Qi Yip, and E. S. Chng, "Low resource language adaptation using two-stage regularization for multilingual asr," in *2024 International Conference on Asian Language Processing (IALP)*, 2024, pp. 332–337.

[76] Y. Lu, J. Yang, and R. Yang, "End-to-end indonesian speech synthesis based on transfer learning and alternate training," in *2021 International Conference on Asian Language Processing (IALP)*, 2021, pp. 30–35.

[77] M. Chen and Z. Yu, "Pre-finetuning for few-shot emotional speech recognition," in *Interspeech 2023*, 2023, pp. 3602–3606.

[78] Z. Zhou, W. Wang, W. Zhang, and Y. Qian, "Exploring effective data utilization for low-resource speech recognition," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8192–8196.

[79] S. Kesiraju, M. SarvaÅ, T. Pavlek, C. Macaire, and A. Ciuba, "Strategies for improving low resource speech to text translation relying on pre-trained asr models," in *Interspeech 2023*, 2023, pp. 2148–2152.

[80] E. van der Westhuizen, H. Kamper, R. Menon, J. Quinn, and T. Niesler, "Feature learning for efficient asr-free keyword spotting in low-resource languages," *Computer Speech & Language*, vol. 71, p. 101275, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000784.

[81] M. Aparna, S. Srivatsa, G. Sai Madhavan, T. B. Dinesh, and S. Srinivasa, "Ai-based assistance for management of oral community knowledge in low-resource and colloquial kannada language," in *Big Data Analytics in Astronomy, Science, and Engineering*, S. Sachdeva and Y. Watanobe, Eds. Cham: Springer Nature Switzerland, 2024, pp. 3–16.

[82] A. Deb, S. Nag, A. Mahapatra, S. Chattopadhyay, A. Marik, P. K. Gayen, S. Sanyal, A. Banerjee, and S. Karmakar, "Beats: Bengali speech acts recognition using multimodal attention fusion," in *Interspeech 2023*, 2023, pp. 3392–3396.

[83] A. Bekarystankyzy, O. Mamyrbayev, and T. Anarbekova, "Integrated end-to-end automatic speech recognition for languages for agglutinative languages," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 6, Jun. 2024. [Online]. Available: https://doi.org/10.1145/3663568.

[84] V. Gupta, "Advances in openasr21 evaluation with increased temporal resolution for speech self-supervised learning models," in *Speech and Computer*, A. Karpov and V. Delić, Eds. Cham: Springer Nature Switzerland, 2025, pp. 69–81.

[85] J. Lee, M. Mimura, and T. Kawahara, "Embedding articulatory constraints for low-resource speech recognition based on large pre-trained model," in *Interspeech 2023*, 2023, pp. 1394–1398.

[86] H. LYU, Z. MA, H. WANG, and X. XIE, "Cnn-ctc based layer transfer model for mongolian speech recognition," *Journal of Chinese Information Processing*, vol. 36, no. 6, pp. 52–60, 2022.

[87] T.-P. Tan, L. Qin, S. Samson Juan, and J. Khaw, "Low resource malay dialect automatic speech recognition modeling using transfer learning from a standard malay model," *Pertanika Journal of Science and Technology*, vol. 32, pp. 1545–1563, 07 2024.

[88] M. A. Kermanshahi, A. Akbari, and B. Nasersharif, "Transfer learning for end-to-end asr to deal with low-resource problem in persian language," in *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, 2021, pp. 1–5.

[89] R. Sahraeian, D. V. Compernolle, and F. de Wet, "Under-resourced speech recognition based on the speech manifold," in *Interspeech 2015*, 2015, pp. 1255–1259.

[90] 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 Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 317–329, 2022.

[91] D. Ng, C. Zhang, R. Zhang, Y. Ma, T. H. Nguyen, C. Ni, S. Zhao, Q. Chen, W. Wang, E. S. Chng, and B. Ma, "Adapter-tuning with effective token-dependent representation shift for automatic speech recognition," in *Interspeech 2023*, 2023, pp. 1319–1323.

[92] X. Li, Y. Wang, X. Liu, K. Su, Z. Li, Y. Wang, B. Jiang, K. Xie, and J. Liu, "Jlms25 and jiao-liao mandarin speech recognition based on multi-dialect knowledge transfer," *Applied Sciences*, vol. 15, no. 3, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/3/1670.

[93] A. Gupta, X. Li, S. K. Rallabandi, and A. W. Black, "Acoustics based intent recognition using discovered phonetic units for low resource languages," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7453–7457.

[94] X. Zheng, C. Zhang, and P. C. Woodland, "Adapting gpt, gpt-2 and bert language models for speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 162–168.

[95] P. Wang, B. BabaAli, and H. Van hamme, "A study into pre-training strategies for spoken language understanding on dysarthric speech," in *Interspeech 2021*, 2021, pp. 36–40.

[96] R. Fan and A. Alwan, "Draft: A novel framework to reduce domain shifting in self-supervised learning and its application to childrens asr," in *Interspeech 2022*, 2022, pp. 4900–4904.

[97] E. Medeiros, L. Corado, L. Rato, P. Quaresma, and P. Salgueiro, "Domain adaptation speech-to-text for low-resource european portuguese using deep learning," *Future Internet*, vol. 15, no. 5, 2023. [Online]. Available: https://www.mdpi.com/1999-5903/15/5/159.

[98] N. M. Pham and V. H. Qu¢n, "Acceleration in state-of-the-art asr applied to a vietnamese transcription system," *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, p. 365372, Jan. 2019. [Online]. Available: https://vjs.ac.vn/index.php/jcc/article/view/13181.

[99] M. Melese, L. Besacier, and M. Meshesha, "Amharic Speech Recognition for Speech Translation," in *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, Paris, France, Jul. 2016, p. n/a. [Online]. Available: https://hal.science/hal-01350050.

[100] A. Abulimiti and T. Schultz, "Automatic speech recognition for Uyghur through multilingual acoustic modeling," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 6444–6449. [Online]. Available: https://aclanthology.org/2020.lrec-1.793/.

[101] C. Yu, Y. Chen, Y. Li, M. Kang, S. Xu, and X. Liu, "Cross-language end-to-end speech recognition research based on transfer learning for the low-resource tujia language," *Symmetry*, vol. 11, no. 2, 2019. [Online]. Available: https://www.mdpi.com/2073-8994/11/2/179.

[102] G. Zellou and M. Lahrouchi, "Linguistic disparities in cross-language automatic speech recognition transfer from arabic to tashlhiyt," *Scientific Reports*, vol. 14, no. 1, p. 313, Jan. 2024.

[103] J. T. Zhou, H. Zhang, D. Jin, and X. Peng, "Dual adversarial transfer for sequence labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 434–446, 2021.

[104] D. Berrebbi, J. Shi, B. Yan, O. Lpez-Francisco, J. Amith, and S. Watanabe, "Combining spectral and self-supervised features for low resource speech recognition and translation," in *Interspeech 2022*, 2022, pp. 3533–3537.

[105] E. Gauthier, L. Besacier, and S. Voisin, "Speed perturbation and vowel duration modeling for asr in hausa and wolof languages," in *Interspeech 2016*, 2016, pp. 3529–3533.

[106] M. U. Farooq and T. Hain, "Investigating the impact of crosslingual acoustic-phonetic similarities on multilingual speech recognition," in *Interspeech 2022*, 2022, pp. 3849–3853.

[107] S. Khurana, "Transfer learning for spoken language processing," Ph.D. dissertation, Massachusetts Institute of Technology, United States-Massachusetts, 2023.

[108] J. Oh, E. Cho, and J.-H. K. and, "Integration of wfst language model in pre-trained korean e2e asr model," *KSII Transactions on Internet and Information Systems*, vol. 18, no. 6, pp. 1692–1705, June 2024.

[109] X. Wu, Y. Jin, J. Wang, Q. Qian, and Y. Guo, "Mkd: Mixup-based knowledge distillation for mandarin end-to-end speech recognition," *Algorithms*, vol. 15, no. 5, 2022. [Online]. Available: https://www.mdpi.com/1999-4893/15/5/160.

[110] D. Bagchi, "Transfer learning approaches for feature denoising and Low-Resource speech recognition," Ph.D. dissertation, The Ohio State University, United States-Ohio, 2020.

[111] H. Wang, S. Wang, W.-Q. Zhang, and J. Bai, "Distilxlsr: A light weight cross-lingual speech representation model," in *Interspeech 2023*, 2023, pp. 2273–2277.

[112] A. Avram, D. Catrina, D. Cercel, M. Dascalu, T. Rebedea, V. F. Pais, and D. Tufis, "Distilling the knowledge of romanian berts using multiple teachers," *CoRR*, vol. abs/2112.12650, 2021. [Online]. Available: https://arxiv.org/abs/2112.12650.

[113] T. Fukuda and S. Thomas, "Knowledge distillation based training of universal asr source models for cross-lingual transfer," in *Interspeech 2021*, 2021, pp. 3450–3454.

[114] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T.-Y. Liu, "Lrspeech: Extremely low-resource speech synthesis and recognition," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 28022812. [Online]. Available: https://doi.org/10.1145/3394486.3403331.

[115] S. Kim, G. Kim, S. Shin, and S. Lee, "Two-stage textual knowledge distillation for end-to-end spoken language understanding," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7463–7467.

[116] A. Seth, S. Ghosh, S. Umesh, and D. Manocha, "Stable distillation: Regularizing continued pre-training for low-resource automatic speech recognition," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10821–10825.

[117] N. T. Le and F. Sadat, "Low-resource machine transliteration using recurrent neural networks of Asian languages," in *Proceedings of the Seventh Named Entities Workshop*, N. Chen, R. E. Banchs, X. Duan, M. Zhang, and H. Li, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 95–100. [Online]. Available: https://aclanthology.org/W18-2414/.

[118] M. Neumann and N. g. Thang Vu, "Cross-lingual and multilingual speech emotion recognition on english and french," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5769–5773.

[119] S. Hara and H. Nishizaki, "Acoustic modeling with a shared phoneme set for multilingual speech recognition without code-switching," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1617–1620.

[120] H. Yadav and S. Sitaram, "A survey of multilingual models for automatic speech recognition," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 5071–5079. [Online]. Available: https://aclanthology.org/2022.lrec-1.542/.

[121] M. Müller, S. Stüker, and A. Waibel, "Language adaptive multilingual ctc speech recognition," in *Speech and Computer*, A. Karpov, R. Potapova, and I. Mporas, Eds. Cham: Springer International Publishing, 2017, pp. 473–482.

[122] J. Billa, "Isi asr system for the low resource speech recognition challenge for indian languages," in *Interspeech 2018*, 2018, pp. 3207–3211.

[123] Y. Hu, S. Settle, and K. Livescu, "Multilingual jointly trained acoustic and written word embeddings," in *Interspeech 2020*, 2020, pp. 1052–1056.

[124] P. Åelasko, L. Moro-Velzquez, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "That sounds familiar: An analysis of phonetic representations transfer across languages," in *Interspeech 2020*, 2020, pp. 3705–3709.

[125] M. Karafit, M. K. Baskar, P. Matjka, K. Vesel, F. Grzl, L. Burget, and J. ernock, "2016 but babel system: Multilingual blstm acoustic model with i-vector based adaptation," in *Interspeech 2017*, 2017, pp. 719–723.

[126] L. Borgholt, T. M. S. Tax, J. D. Havtorn, L. Maale, and C. Igel, "On scaling contrastive representations for low-resource speech recognition," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3885–3889.

[127] Z. Kozhirbayev, "Leveraging wav2vec2.0 for kazakh speech recognition: An experimental study," in *Computational Science and Its Applications-ICCSA 2024*, O. Gervasi, B. Murgante, C. Garau, D. Taniar, A. M. A. C. Rocha, and M. N. Faginas Lago, Eds. Cham: Springer Nature Switzerland, 2024, pp. 120–132.

[128] Q. shi Zhu, J. Zhang, M. hui Wu, X. Fang, and L.-R. Dai, "An improved wav2vec 2.0 pre-training approach using enhanced local dependency modeling for speech recognition," in *Interspeech 2021*, 2021, pp. 4334–4338.

[129] P. Åelasko, S. Feng, L. Moro Velzquez, 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230822000067.

[130] Y. Shi, Z. Tang, L. Lit, Z. Zhang, and D. Wang, "Map and relabel: Towards almost-zero resource speech recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 591–595.

[131] S. Feng and T. Lee, "Exploiting cross-lingual speaker and phonetic diversity for unsupervised subword modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2000–2011, 2019.

[132] S. Khurana, A. Laurent, and J. Glass, "Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6647–6651.

[133] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6419–6423.

[134] Z.-Q. Zhang, Y. Song, M.-H. Wu, X. Fang, I. McLoughlin, and L.-R. Dai, "Cross-Lingual self-training to learn multilingual representation for Low-Resource speech recognition," *Circuits, Systems, and Signal Processing*, vol. 41, no. 12, pp. 6827–6843, Dec. 2022.

[135] B. Abraham, T. Seeram, and S. Umesh, "Transfer learning and distillation techniques to improve the acoustic modeling of low resource languages," in *Interspeech 2017*, 2017, pp. 2158–2162.

[136] K. Choi and H.-M. Park, "Distilling a pretrained language model to a multilingual asr model," in *Interspeech 2022*, 2022, pp. 2203–2207.

[137] B. Song, W. Zhang, M. Cai, J. Liu, and M. T. Johnson, "Query-by-example spoken term detection based on phonetic posteriorgram," in *Proceedings of the 2015 InternationalConference on Education, Management and Computing Technology*. Atlantis Press, 2015/06, pp. 1251–1256. [Online]. Available: https://doi.org/10.2991/icemct-15.2015.256.

[138] J. Badenhorst, L. Martinus, and F. de Wet, "Blstm harvesting of auxiliary nchlt speech data," in *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 2019, pp. 123–128.

[139] Y. Xie, R. Liang, X. Zhao, Z. Liang, and J. Du, "Weighted gradient pretrain for low-resource speech emotion recognition," *IEICE Trans. Inf. Syst.*, vol. 105-D, pp. 1352–1355, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250185259.

[140] S. Dalmia, X. Li, F. Metze, and A. W. Black, "Domain robust feature extraction for rapid low resource asr development," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 258–265.

[141] Q. Zhan, X. Xie, C. Hu, J. Zuluaga-Gomez, J. Wang, and H. Cheng, "Domain-adversarial based model with phonological knowledge for cross-lingual speech recognition," *Electronics*, vol. 10, no. 24, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/24/3172.

[142] Z.-T. Liu, B.-H. Wu, M.-T. Han, W.-H. Cao, and M. Wu, "Speech emotion recognition based on meta-transfer learning with domain adaption," *Applied Soft Computing*, vol. 147, p. 110766, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623007846.

[143] G. Chatzoudis, M. Plitsis, S. Stamouli, A. Dimou, N. Katsamanis, and V. Katsouros, "Zero-shot cross-lingual aphasia detection using automatic speech recognition," in *Interspeech 2022*, 2022, pp. 2178–2182.

[144] H. Yen, P.-J. Ku, C.-H. H. Yang, H. Hu, S. M. Siniscalchi, P.-Y. Chen, and Y. Tsao, "Neural model reprogramming with similarity based mapping for low-resource spoken command recognition," in *Interspeech 2023*, 2023, pp. 3317–3321.

[145] V. N. Sukhadia and S. Umesh, "Domain adaptation of low-resource target-domain models using well-trained asr conformer models," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 295–301.

[146] P. Wu, J. Shi, Y. Zhong, S. Watanabe, and A. W. Black, "Cross-lingual transfer for speech processing using acoustic language similarity," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1050–1057.

[147] Z. Li, P. Blumenberg, J. Liu, T. Graave, T. Lohrenz, S. Kunzmann, and T. Fingscheidt, "Interleaved audio/audiovisual transfer learning for av-asr in low-resourced languages," in *Interspeech 2024*, 2024, pp. 2524–2528.

[148] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual mlp features for low-resource lvcsr systems," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4269–4272.

[149] R. Sahraeian, D. Van Compernolle, and F. de Wet, "Using generalized maxout networks and phoneme mapping for low resource asr- a case study on flemish-afrikaans," in *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2015, pp. 112–117.

[150] R. Sahraeian and D. Van Compernolle, "A study of rank-constrained multilingual dnns for low-resource asr," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5420–5424.

[151] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pretraining for low-resource speech-to-text translation," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7909–7913.

[152] M. Jakubec, E. Lieskovska, R. Jarina, M. Spisiak, and P. Kasak, "Speech emotion recognition using transfer learning: Integration of advanced speaker embeddings and image recognition models," *Applied Sciences*, vol. 14, no. 21, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/21/9981.

[153] D. Liu, J. Xu, P. Zhang, and Y. Yan, "Investigation of knowledge transfer approaches to improve the acoustic modeling of vietnamese asr system," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 5, pp. 1187–1195, 2019.

[154] A. Zhang and C. Ni, "Research on low-resource mongolian speech recognition," *Computer Science*, vol. 44, pp. 318–322, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:212749123.

[155] T. Alume, S. Tsakalidis, and R. Schwartz, "Improved multilingual training of stacked neural network acoustic models for low resource languages," in *Interspeech 2016*, 2016, pp. 3883–3887.

[156] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafit, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 521–527.

[157] K. Fatehi, M. Torres Torres, and A. Kucukyilmaz, "Scoutwav: Two-step fine-tuning on self-supervised automatic speech recognition for low-resource environments," in *Interspeech 2022*, 2022, pp. 3523–3527.

[158] S. Feng, M. Tu, R. Xia, C. Huang, and Y. Wang, "Language-universal phonetic encoder for low-resource speech recognition," in *Interspeech 2023*, 2023, pp. 1429–1433.

[159] T. Wullach and S. E. Chazan, "Dont be so sure! boosting asr decoding via confidence relaxation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 13780–13788, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26614.

[160] S. Zaiem, R. Algayres, T. Parcollet, S. Essid, and M. Ravanelli, "Fine-tuning strategies for faster inference using speech self-supervised models: A comparative study," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023, pp. 1–5.

[161] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, "Dysarthric speech recognition based on deep metric learning," in *Interspeech 2020*, 2020, pp. 4796–4800.

[162] J. Hellarawa and U. Thayasivam, "Domain specific intent classification with bilstm," in *2022 International Conference on Asian Language Processing (IALP)*, 2022, pp. 265–270.

[163] J. Yang, J. Wei, K. Khysru, J. Xu, W. Lu, W. Ke, and X. Yang, "Effective fine-tuning method for tibetan low-resource dialect speech recognition," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 960–965.

[164] N. Tomashenko, A. Caubrire, and Y. Estve, "Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech," in *Interspeech 2019*, 2019, pp. 824–828.

[165] C. Wang, J. Pino, and J. Gu, "Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation," in *Interspeech 2020*, 2020, pp. 4731–4735.

[166] C. Yi, S. Zhou, and B. Xu, "Efficiently fusing pretrained acoustic and linguistic encoders for low-resource speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 788–792, 2021.

[167] Z. Zeng, V. T. Pham, H. Xu, Y. Khassanov, E. S. Chng, C. Ni, and B. Ma, "Leveraging text data using hybrid transformer-lstm based end-to-end asr in transfer learning," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[168] Y. Qian and Z. Zhou, "Optimizing data usage for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 394–403, 2022.

[169] H. Yadav, A. Gupta, S. K. Rallabandi, A. W. Black, and R. R. Shah, "Intent classification using pre-trained language agnostic embeddings for low resource languages," in *Interspeech 2022*, 2022, pp. 3473–3477.

[170] N. Karpov, A. Denisenko, and F. Minkin, "Golos: Russian dataset for speech research," in *Interspeech 2021*, 2021, pp. 1419–1423.

[171] S. Leivaditi, T. Matsushima, M. Coler, S. Nayak, and V. Verkhodanova, "Fine-tuning strategies for dutch dysarthric speech recognition: Evaluating the impact of healthy, disease-specific, and speaker-specific data," in *Interspeech 2024*, 2024, pp. 1295–1299.

[172] T. Alume and J. Kong, "Combining hybrid and end-to-end approaches for the openasr20 challenge," in *Interspeech 2021*, 2021, pp. 4349–4353.

[173] T. Deng, J. Wei, J. Yang, M. Guo, W. Ke, X. Yang, and W. Lu, "Breaking the corpus bottleneck for multi-dialect speech recognition with flexible adapters," in *Artificial Neural Networks and Machine Learning-ICANN 2024*, M. Wand, K. Malinovská, J. Schmidhuber, and I. V. Tetko, Eds. Cham: Springer Nature Switzerland, 2024, pp. 3–15.

[174] D. Kulshreshtha, S. Dingliwal, B. Houston, and S. Bodapati, "Multilingual contextual adapters to improve custom word recognition in low-resource languages," in *Interspeech 2023*, 2023, pp. 3302–3306.

[175] G. I. Winata, G. Wang, C. Xiong, and S. Hoi, "Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition," in *Interspeech 2021*, 2021, pp. 2451–2455.

[176] S. Udupa, J. Bandekar, S. Kumar, D. G, S. B, A. S, S. Murthy, P. Pai, S. Raghavan, R. Nanavati, and P. K. Ghosh, "Adapter pre-training for improved speech recognition in unseen domains using low resource adapter tuning of self-supervised models," in *Interspeech 2024*, 2024, pp. 2529–2533.

[177] Y. Liu, X. Yang, J. Zhang, Y. Xi, and D. Qu, "Taml-adapter: Enhancing adapter tuning through task-agnostic meta-learning for low-resource automatic speech recognition," *IEEE Signal Processing Letters*, vol. 32, pp. 636–640, 2025.

[178] X. Geng, T. Xu, K. Wei, B. Mu, H. Xue, H. Wang, Y. Li, P. Guo, Y. Dai, L. Li, M. Shao, and L. Xie, "Unveiling the potential of llm-based asr on chinese open-source datasets," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2024, pp. 26–30.

[179] M. Wiesner, O. Adams, D. Yarowsky, J. Trmal, and S. Khudanpur, "Zero-shot pronunciation lexicons for cross-language acoustic model transfer," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1048–1054.

[180] K. Wongpatikaseree, S. Singkul, N. Hnoohom, and S. Yuenyong, "Real-time end-to-end speech emotion recognition with cross-domain adaptation," *Big Data and Cognitive Computing*, vol. 6, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2504-2289/6/3/79.

[181] A. Johnson, "Towards inclusive Low-Resource speech technologies: A case study of educational systems for african american English-Speaking children," Ph.D. dissertation, University of California, Los Angeles, United States-California, 2024.

[182] J. Zuluaga-Gomez, A. Prasad, I. Nigmatulina, S. S. Sarfjoo, P. Motlicek, M. Kleinert, H. Helmke, O. Ohneiser, and Q. Zhan, "How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 205–212.

[183] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "Multilingual acoustic and language modeling for ethio-semitic languages," in *Interspeech 2020*, 2020, pp. 1047–1051.

[184] S. Li, C. Ding, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "End-to-end articulatory attribute modeling for low-resource multilingual speech recognition," in *Interspeech 2019*, 2019, pp. 2145–2149.

[185] A. Bekarystankyzy, O. Mamyrbayev, M. Mendes, A. Fazylzhanova, and M. Assam, "Multilingual end-to-end ASR for low-resource turkic languages with common alphabets," *Scientific Reports*, vol. 14, no. 1, p. 13835, Jun. 2024.

[186] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8249–8253.

[187] A. Ullah, A. Ragano, and A. Hines, "Reduce, reuse, recycle: Is perturbed data better than other language augmentation for low resource self-supervised speech models," in *Interspeech 2024*, 2024, pp. 77–81.

[188] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estve, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, "Lebenchmark: A reproducible framework for assessing selfsupervised representation learning from speech," in *Interspeech 2021*, 2021, pp. 1439–1443.

[189] T. Javed, S. Doddapaneni, A. Raman, K. S. Bhogale, G. Ramesh, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Towards building asr systems for the next billion users," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10813–10821, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21327.

[190] R. Huang and B. Mak, "wav2vec 2.0 asr for cantonese-speaking older adults in a clinical setting," in *Interspeech 2023*, 2023, pp. 4958–4962.

[191] P. Hemant and M. Narvekar, "Development of a code-switched hindi-marathi dataset and transformer-based architecture for enhanced speech recognition using dynamic switching algorithms," *Applied Acoustics*, vol. 230, p. 110408, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X24005590.

[192] A. Gupta, A. Jayasimha, A. Maghan, S. Singh, D. Gowda, and C. Kim, "Voice to action: Spoken language understanding for memory-constrained systems," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 473–479.

[193] R. Wang, Q. Bai, J. Ao, L. Zhou, Z. Xiong, Z. Wei, Y. Zhang, T. Ko, and H. Li, "Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit bert," in *Interspeech 2022*, 2022, pp. 1686–1690.

[194] E. G. Garcia, E. Mengusoglu, and E. Janke, "Multilingual acoustic models for speech recognition in low-resource devices," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP '07*, vol. 4, 2007, pp. IV–981–IV–984.

[195] Y. Qian, J. Xu, and J. Liu, "Multi-stream posterior features and combining subspace gmms for low resource lvcsr," *Chinese Journal of Electronics*, vol. 22, no. 2, pp. 291–295, 2013.

[196] S. Liu and K. C. Sim, "Multi-stream temporally varying weight regression for cross-lingual speech recognition," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 434–439.

[197] S. Zhou, Y. Zhao, S. Xu, and B. Xu, "Multilingual recurrent neural networks with residual learning for low-resource speech recognition," in *Interspeech 2017*, 2017, pp. 704–708.

[198] B. Sen, A. Agarwal, M. Sai Ganesh, and A. Kumar Vuppala, "Reed: An approach towards quickly bootstrapping multilingual acoustic models," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 272–279.

[199] M. U. Farooq, D. A. H. Narayana, and T. Hain, "Non-linear pairwise language mappings for low-resource multilingual acoustic model fusion," in *Interspeech 2022*, 2022, pp. 4850–4854.

[200] S. Thomas, K. Audhkhasi, J. Cui, B. Kingsbury, and B. Ramabhadran, "Multilingual data selection for low resource speech recognition," in *Interspeech 2016*, 2016, pp. 3853–3857.

[201] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nussbaum-Thom, M. Picheny, Z. Tske, P. Golik, R. Schlter, H. Ney, M. J. F. Gales, K. M. Knill, A. Ragni, H. Wang, and P. Woodland, "Multilingual representations for low resource speech recognition and keyword search," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 259–266.

[202] T. Sercu, G. Saon, J. Cui, X. Cui, B. Ramabhadran, B. Kingsbury, and A. Sethy, "Network architectures for multilingual speech representation learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5295–5299.

[203] M. Wiesner, M. Sarma, A. Arora, D. Raj, D. Gao, R. Huang, S. Preet, M. Johnson, Z. Iqbal, N. Goel, J. Trmal, L. P. G. Perera, and S. Khudanpur, "Training hybrid models on noisy transliterated transcripts for code-switched speech recognition," in *Interspeech 2021*, 2021, pp. 2906–2910.

[204] X. Li, S. Dalmia, A. W. Black, and F. Metze, "Multilingual speech recognition with corpus relatedness sampling," in *Interspeech 2019*, 2019, pp. 2120–2124.

[205] C. Zhu, K. An, H. Zheng, and Z. Ou, "Multilingual and crosslingual speech recognition using phonological-vector based phone embeddings," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1034–1041.

[206] N. Gaur, T. Chen, E. Variani, P. Haghani, B. Ramabhadran, and P. J. Moreno, "Multilingual second-pass rescoring for automatic speech recognition systems," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6407–6411.

[207] W. Hou, Y. Wang, S. Gao, and T. Shinozaki, "Meta-adapter: Efficient cross-lingual adaptation with meta-learning," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7028–7032.

[208] Y. Zhu, P. Haghani, A. Tripathi, B. Ramabhadran, B. Farris, H. Xu, H. Lu, H. Sak, I. Leal, N. Gaur, P. J. Moreno, and Q. Zhang, "Multilingual speech recognition with self-attention structured parameterization," in *Interspeech 2020*, 2020, pp. 4741–4745.

[209] X. Yue, X. Gao, X. Qian, and H. Li, "Adapting pre-trained self-supervised learning model for speech recognition with light-weight adapters," *Electronics*, vol. 13, no. 1, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/1/190.

[210] Q. Hu, Y. Zhang, X. Zhang, Z. Han, and X. Yu, "Cam: A cross-lingual adaptation framework for low-resource language speech recognition," *Information Fusion*, vol. 111, p. 102506, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253524002847.

[211] A. Ansell, E. M. Ponti, J. Pfeiffer, S. Ruder, G. Glavaš, I. Vulić, and A. Korhonen, "MAD-G: Multilingual adapter generation for efficient cross-lingual transfer," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4762–4781. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.410/.

[212] G. Zheng, Y. Xiao, K. Gong, P. Zhou, X. Liang, and L. Lin, "Wav-BERT: Cooperative acoustic and linguistic representation learning for low-resource speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2765–2777. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.236/.

[213] Z. Yu, Y. Zhang, K. Qian, C. Wan, Y. Fu, Y. Zhang, and Y. C. Lin, "Master-ASR: Achieving multilingual scalability and low-resource adaptation in ASR with modular learning," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 40475–40487. [Online]. Available: https://proceedings.mlr.press/v202/yu23l.html.

[214] Y. Fu, Y. Zhang, K. Qian, Z. Ye, Z. Yu, C.-I. J. Lai, and C. Lin, "Losses can be blessings: Routing self-supervised speech representations towards efficient multilingual and multitask speech processing," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 20902–20920. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/83d349b6eb8125588b5f091e2d47525c-Paper-Conference.pdf.

[215] Y. Ma, T. H. Nguyen, and B. Ma, "Cpt: Cross-modal prefix-tuning for speech-to-text translation," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6217–6221.

[216] Y. Chen, W. Zhang, H. Zhang, X. Yang, and D. Qu, "Exploring the potential of prompting methods in low-resource speech recognition with whisper," in *Natural Language Processing and Chinese Computing*, D. F. Wong, Z. Wei, and M. Yang, Eds. Singapore: Springer Nature Singapore, 2025, pp. 382–393.

[217] S. Feng, P. Åelasko, L. Moro-Velzquez, and O. Scharenborg, "Unsupervised acoustic unit discovery by leveraging a language-independent subword discriminative feature representation," in *Interspeech 2021*, 2021, pp. 1534–1538.

[218] M. Doumbouya, L. Einstein, and C. Piech, "Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 14757–14765, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17733.

[219] J. Kim, M. Kumar, D. Gowda, A. Garg, and C. Kim, "Semi-supervised transfer learning for language expansion of end-to-end speech recognition models to low-resource languages," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 984–988.

[220] A. Pasad, F. Wu, S. Shon, K. Livescu, and K. Han, "On the use of external data for spoken named entity recognition," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 724–737. [Online]. Available: https://aclanthology.org/2022.naacl-main.53/.

[221] C. Lim, "Towards deep learning on speech recognition for khmer language," Ph.D. dissertation, University of Missouri-Columbia, United States-Missouri, 2016.

[222] M. S. Mary N J, V. M. Shetty, and S. Umesh, "Investigation of methods to improve the recognition performance of tamil-english code-switched data in transformer framework," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7889–7893.

[223] H. Xu, V. H. Do, X. Xiao, and E. S. Chng, "A comparative study of bnf and dnn multilingual training on cross-lingual low-resource speech recognition," in *Interspeech 2015*, 2015, pp. 2132–2136.

[224] P. Pandey, P. Hebbar, P. Borole, S. Satpal, R. Motlani, R. Mehta, N. Nayak, and R. Srikanth, "Towards bootstrapping acoustic models for resource poor indian languages," in *2017 Twenty-third National Conference on Communications (NCC)*, 2017, pp. 1–4.

[225] L. Maison and Y. Estve, "Some voices are too common: Building fair speech recognition systems using the commonvoice dataset," in *Interspeech 2023*, 2023, pp. 4428–4432.

[226] H. Palivela, M. Narvekar, D. Asirvatham, S. Bhushan, V. Rishiwal, and U. Agarwal, "Code-switching asr for low-resource indic languages: A hindi-marathi case study," *IEEE Access*, vol. 13, pp. 9171–9198, 2025.

[227] A. S. Sundar, C.-H. H. Yang, D. M. Chan, S. Ghosh, V. Ravichandran, and P. S. Nidadavolu, "Multimodal attention merging for improved speech recognition and audio event classification," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 655–659.

[228] T. Pekarek Rosin and S. Wermter, "Replay to remember: Continual layer-specific fine-tuning for german speech recognition," in *Artificial Neural Networks and Machine Learning-ICANN 2023*, L. Iliadis, A. Papaleonidas, P. Angelov, and C. Jayne, Eds. Cham: Springer Nature Switzerland, 2023, pp. 489–500.

[229] P. Do, M. Coler, J. Dkstra, and E. Klabbers, "The effects of input type and pronunciation dictionary usage in transfer learning for low-resource text-to-speech," in *Interspeech 2023*, 2023, pp. 5461–5465.

[230] B. Pulugundla, M. K. Baskar, S. Kesiraju, E. Egorova, M. Karafit, L. Burget, and J. ernock, "But system for low resource indian language asr," in *Interspeech 2018*, 2018, pp. 3182–3186.

[231] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 58–68. [Online]. Available: https://aclanthology.org/N19-1006/.

[232] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end asr with language model fusion," in *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6096–6100.

[233] B. Thai, R. Jimerson, D. Arcoraci, E. Prud'hommeaux, and R. Ptucha, "Synthetic data augmentation for improving low-resource asr," in *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, 2019, pp. 1–9.

[234] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Interspeech 2021*, 2021, pp. 1509–1513.

[235] S. Khare, A. Mittal, A. Diwan, S. Sarawagi, P. Jyothi, and S. Bharadwaj, "Low resource asr: The surprising effectiveness of high resource transliteration," in *Interspeech 2021*, 2021, pp. 1529–1533.

[236] K. Siminyu, X. Li, A. Anastasopoulos, D. R. Mortensen, M. R. Marlo, and G. Neubig, "Phoneme recognition through fine tuning of phonetic representations: A case study on luhya language varieties," in *Interspeech 2021*, 2021, pp. 271–275.

[237] A. Rouhe, A. Virkkunen, J. Leinonen, and M. Kurimo, "Low resource comparison of attention-based and hybrid asr exploiting wav2vec 2.0," in *Interspeech 2022*, 2022, pp. 3543–3547.

[238] N. Abdou Mohamed, A. Allak, K. Gaanoun, I. Benelallam, Z. Erraji, and A. Bahafid, "Multilingual speech recognition initiative for african languages," *International Journal of Data Science and Analytics*, Nov. 2024.

[239] T. Alume, K. Kukk, V.-B. Le, C. Barras, A. Messaoudi, and W. Ben Kheder, "Exploring the impact of pretrained models and web-scraped data for the 2022 nist language recognition evaluation," in *Interspeech 2023*, 2023, pp. 516–520.

[240] A. Pieiro-Martn, C. Garca-Mateo, L. Docio-Fernandez, M. del Carmen Lpez-Prez, and G. Rehm, "Weighted cross-entropy for low-resource languages in multilingual speech recognition," in *Interspeech 2024*, 2024, pp. 1235–1239.

[241] L. Wang, S. Feng, M. Hasegawa-Johnson, and C. Yoo, "Self-supervised semantic-driven phoneme discovery for zero-resource speech recognition," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8027–8047. [Online]. Available: https://aclanthology.org/2022.acl-long.553/.

[242] K. D. N, P. Wang, and B. Bozza, "Using large self-supervised models for low-resource speech recognition," in *Interspeech 2021*, 2021, pp. 2436–2440.

[243] H. Le, F. Barbier, H. Nguyen, N. Tomashenko, S. Mdhaffar, S. G. Gahbiche, B. Lecouteux, D. Schwab, and Y. Estève, "ON-TRAC' systems for the IWSLT 2021 low-resource speech translation and multilingual speech translation shared tasks," in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, M. Federico, A. Waibel, M. R. Costa-jussà, J. Niehues, S. Stuker, and E. Salesky, Eds. Bangkok, Thailand (online): Association for Computational Linguistics, Aug. 2021, pp. 169–174. [Online]. Available: https://aclanthology.org/2021.iwslt-1.20/.

[244] K. Kurihara and M. Sano, "Low-resourced phonetic and prosodic feature estimation with self-supervised-learning-based acoustic modeling," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 640–644.

[245] Y. Shen, Q. Liu, Z. Fan, J. Liu, and A. Wumaier, "Self-supervised pre-trained speech representation based end-to-end mispronunciation detection and diagnosis of mandarin," *IEEE Access*, vol. 10, pp. 106451–106462, 2022.

[246] A. K. Kandji, C. Ba, and S. Ndiaye, "State-of-the-art review on recent trends in automatic speech recognition," in *Emerging Technologies for Developing Countries*, M. Masinde, S. Mübs, and A. Bagula, Eds. Cham: Springer Nature Switzerland, 2024, pp. 185–203.

[247] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, P. Xu, and P. Fung, "Meta-transfer learning for code-switched speech recognition," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 3770–3776. [Online]. Available: https://aclanthology.org/2020.acl-main.348/.

[248] G. I. Winata, "Multilingual transfer learning for Code-Switched language and speech neural modeling," Ph.D. dissertation, Hong Kong University of Science and Technology (Hong Kong), Hong Kong, 2021.

[249] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7844–7848.

[250] Y. Xiao, K. Gong, P. Zhou, G. Zheng, X. Liang, and L. Lin, "Adversarial meta sampling for multilingual low-resource speech recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14112–14120, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17661.

[251] Y. Chen, H. Zhang, X. Yang, W. Zhang, and D. Qu, "Improving cross-lingual low-resource speech recognition by task-based meta polyloss," *Computer Speech & Language*, vol. 87, p. 101648, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230824000317.

[252] S. Singh, R. Wang, and F. Hou, "Improved meta learning for low resource speech recognition," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4798–4802.

[253] Y. Chen, X. Yang, H. Zhang, W. Zhang, D. Qu, and C. Chen, "Meta adversarial learning improves low-resource speech recognition," *Computer Speech & Language*, vol. 84, p. 101576, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230823000955.

[254] Y. Zhu, R. Fan, and A. Alwan, "Towards better meta-initialization with task augmentation for kindergarten-aged speech recognition," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8582–8586.

[255] C. S. Anoop and A. G. Ramakrishnan, "Meta-learning for indian languages: Performance analysis and improvements with linguistic similarity measures," *IEEE Access*, vol. 11, pp. 82050–82064, 2023.

[256] R. Zhou, A. Ito, and T. Nose, "Character expressions in meta-learning for extremely low resource language speech recognition," in *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, ser. ICMLC '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 525529. [Online]. Available: https://doi.org/10.1145/3651671.3651730.

[257] S. Qin, L. Wang, S. Li, Y. Lin, and J. Dang, "Finer-grained modeling units-based meta-learning for low-resource tibetan speech recognition," in *Interspeech 2022*, 2022, pp. 2133–2137.

[258] Y. Chen, T. Niu, H. Zhang, W. Zhang, and D. Qu, "Meta-prompt: Boosting whisper's performance in low-resource speech recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 3039–3043, 2024.

[259] D. Eledath, A. Baby, and S. Singh, "Robust speech recognition using meta-learning for low-resource accents," in *2024 National Conference on Communications (NCC)*, 2024, pp. 1–6.

[260] R. Zhou, T. Koshikawa, A. Ito, T. Nose, and C.-P. Chen, "Multilingual meta-transfer learning for low-resource speech recognition," *IEEE Access*, vol. 12, pp. 158493–158504, 2024.

[261] S. Chopra, P. Mathur, R. Sawhney, and R. R. Shah, "Meta-learning for low-resource speech emotion recognition," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6259–6263.

[262] D. Eledath and V. Ramasubramanian, "Few-shot learning for low-resource e2e asr: Mono-, cross- and multi-lingual scenarios," in *2024 International Conference on Signal Processing and Communications (SPCOM)*, 2024, pp. 1–5.

[263] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and P. Fung, "Learning fast adaptation on cross-accented speech recognition," in *Interspeech 2020*, 2020, pp. 1276–1280.

[264] Y. Gao, J. Feng, C. Deng, and S. Zhang, "Meta auxiliary learning for low-resource spoken language understanding," in *Interspeech 2022*, 2022, pp. 2703–2707.

[265] B. M. Lake and M. Baroni, "Human-like systematic generalization through a meta-learning neural network," *Nature*, vol. 623, no. 7985, pp. 115–121, Nov. 2023.

[266] T. G. Fantaye, J. Yu, and T. T. Hailu, "Investigation of various hybrid acoustic modeling units via a multitask learning and deep neural network technique for lvcsr of the low-resource language, amharic," *IEEE Access*, vol. 7, pp. 105593–105608, 2019.

[267] D. Chen and B. K.-W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1172–1183, 2015.

[268] J. Bai, B. Li, Y. Zhang, A. Bapna, N. Siddhartha, K. Chai Sim, and T. N. Sainath, "Joint unsupervised and supervised training for multilingual asr," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6402–6406.

[269] N. Yolwas and W. Meng, "Jsum: A multitask learning speech recognition model for jointly supervised and unsupervised learning," *Applied Sciences*, vol. 13, no. 9, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5239.

[270] 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.

[271] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinozaki, "Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning," in *Interspeech 2020*, 2020, pp. 1037–1041.

[272] S. Madikeri, P. Motlicek, and H. Bourlard, "Multitask adaptation with lattice-free mmi for multi-genre speech recognition of low resource languages," in *Interspeech 2021*, 2021, pp. 4329–4333.

[273] S. Miwa and A. Kai, "Dialect speech recognition modeling using corpus of japanese dialects and self-supervised learning-based model xlsr," in *Interspeech 2023*, 2023, pp. 4928–4932.

[274] Z. Dan, Y. Zhao, X. Bi, L. Wu, and Q. Ji, "Multi-task learning with auxiliary cross-attention transformer for low-resource multi-dialect speech recognition," in *Natural Language Processing and Chinese Computing*, W. Lu, S. Huang, Y. Hong, and X. Zhou, Eds. Cham: Springer International Publishing, 2022, pp. 107–118.

[275] H. Wang, F. Gao, Y. Zhao, L. Yang, J. Yue, and H. Ma, "Multitask learning with local attention for tibetan speech recognition," *Complexity*, vol. 2020, no. 1, p. 8894566, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/8894566.

[276] O. Mamyrbayev, S. Pavlov, A. Bekarystankyzy, D. Oralbekova, B. Zhumazhanov, L. Azarova, D. Mussayeva, T. Koval, K. Gromaszek, N. Issimov, and K. Shiyapov, "Neurorecognition visualization in multitask end-to-end speech," in *Optical Fibers and Their Applications 2023*, W. Wójcik, Z. Omiotek, and A. Smolarz, Eds., vol. 12985, International Society for Optics and Photonics. SPIE, 2023, p. 129850G. [Online]. Available: https://doi.org/10.1117/12.3022727.

[277] W. Chen, "Modeling phones, keywords, topics and intents in spoken languages," Ph.D. dissertation, University of Illinois at Urbana-Champaign, United States-Illinois, 2019.

[278] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Multitask feature learning for low-resource query-by-example spoken term detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1329–1339, 2017.

[279] Y. Jianbin, K. Jian, Z. Wei-Qiang, and L. Jia, "Mutitask learning based muti-examples keywords spotting in low resource condition," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, 2018, pp. 581–585.

[280] M. M. E. Noor, Y.-J. Lu, S.-S. Wang, S. Ghose, C.-Y. Chang, R. E. Zezario, S. Ahmed, W.-H. Chung, Y. Tsao, and H.-M. Wang, "Investigation of a single-channel frequency-domain speech enhancement network to improve end-to-end bengali automatic speech recognition under unseen noisy conditions," in *2021 24th*

*Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2021, pp. 7–12.

[281] P. Swietojanski, J. Li, and J.-T. Huang, "Investigation of maxout networks for speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7649–7653.

[282] Q. Meeus, M. Francine Moens, and H. Van hamme, "Multitask learning for low resource spoken language understanding," in *Interspeech 2022*, 2022, pp. 4073–4077.

[283] L. Bansal, S. P. Dubagunta, M. Chetlur, P. Jagtap, and A. Ganapathiraju, "On the efficacy and noise-robustness of jointly learned speech emotion and automatic speech recognition," in *Interspeech 2023*, 2023, pp. 1863–1867.

[284] Y. Liu, Y. Zhao, X. Xu, L. Xu, and X. Zhang, "Multi-task pre-training for lhasa-tibetan speech recognition," in *Artificial Neural Networks and Machine Learning-ICANN 2023*, L. Iliadis, A. Papaleonidas, P. Angelov, and C. Jayne, Eds. Cham: Springer Nature Switzerland, 2023, pp. 78–90.

[285] D. Eledath, N. R. Thurlapati, V. Pavithra, T. Banerjee, and V. Ramasubramanian, "Few-shot learning for e2e speech recognition: architectural variants for support set generation," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 444–448.

[286] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, A. Mittal, P. K. Ghosh, P. Jyothi, K. Bali, V. Seshadri, S. Sitaram, S. Bharadwaj, J. Nanavati, R. Nanavati, and K. Sankaranarayanan, "Mucs 2021: Multilingual and code-switching asr challenges for low resource indian languages," in *Interspeech 2021*, 2021, pp. 2446–2450.

[287] J. Meyer, "Multi-Task and transfer learning in Low-Resource speech recognition," Ph.D. dissertation, The University of Arizona, United States-Arizona, 2019.

[288] P. Denisov, M. Mager, and N. T. Vu, "IMS' systems for the IWSLT 2021 low-resource speech translation task," in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, M. Federico, A. Waibel, M. R. Costa-jussà, J. Niehues, S. Stuker, and E. Salesky, Eds. Bangkok, Thailand (online): Association for Computational Linguistics, Aug. 2021, pp. 175–181. [Online]. Available: https://aclanthology.org/2021.iwslt-1.21/.

[289] T. A. Dinh, D. Liu, and J. Niehues, "Tackling data scarcity in speech translation using zero-shot multilingual machine translation techniques," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6222–6226.

[290] Z. Huang, J. Xu, Q. Zhao, and P. Zhang, "A two-fold cross-validation training framework combined with meta-learning for code-switching speech recognition," *IEICE Trans. Inf. Syst.*, vol. 105-D, pp. 1639–1642, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251990391.

[291] X. Zeng, L. Li, and Q. Liu, "Multilingual speech translation with unified transformer: Huawei Noah's ark lab at IWSLT 2021," in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, M. Federico, A. Waibel, M. R. Costa-jussà, J. Niehues, S. Stuker, and E. Salesky, Eds. Bangkok, Thailand (online): Association for Computational Linguistics, Aug. 2021, pp. 149–153. [Online]. Available: https://aclanthology.org/2021.iwslt-1.17/.

[292] W. Yu, C. Tang, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Connecting speech encoder and large language model for asr," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12637–12641.

[293] Z. Chen, H. Huang, O. Hrinchuk, K. C. Puvvada, N. R. Koluguri, P. Åelasko, J. Balam, and B. Ginsburg, "Bestow: Efficient and streamable speech language model with the best of two worlds in gpt and t5," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 147–154.

[294] K. Hu, T. N. Sainath, B. Li, Y. Zhang, Y. Cheng, T. Wang, Y. Zhang, and F. Liu, "Improving multilingual and code-switching asr using large language model generated text," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[295] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5723–5738. [Online]. Available: https://aclanthology.org/2022.acl-long.393/.

[296] C.-J. Hsu, H.-L. Chung, H.-Y. Lee, and Y. Tsao, "T5lephone: Bridging speech and text self-supervised models for spoken language understanding via phoneme level t5," in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[297] K. Deng, J. Guo, Y. Ma, N. Moritz, P. C. Woodland, O. Kalinli, and M. Seltzer, "Transducer-llama: Integrating llms into streamable transducer-based speech recognition," *CoRR*, vol. abs/2412.16464, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2412.16464.

[298] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, "Llama-omni: Seamless speech interaction with large language models," 2025. [Online]. Available: https://arxiv.org/abs/2409.06666

[299] S.-E. Kim, B. R. Chernyak, O. Seleznova, J. Keshet, M. Goldrick, and A. R. Bradlow, "Automatic recognition of second language speech-in-noise," *JASA Express Letters*, vol. 4, no. 2, p. 025204, 022024. [Online]. Available: https://doi.org/10.1121/10.0024877.

[300] N. Phan, A. von Zansen, M. Kautonen, E. Voskoboinik, T. Grosz, R. Hilden, and M. Kurimo, "Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task," in *Interspeech 2024*, 2024, pp. 317–321.

[301] B. Talafha, A. Waheed, and M. Abdul-Mageed, "N-shot benchmarking of whisper on diverse arabic speech recognition," in *Interspeech 2023*, 2023, pp. 5092–5096.

[302] P. Peng, B. Yan, S. Watanabe, and D. Harwath, "Prompting the hidden talent of web-scale speech models for zero-shot task generalization," in *Interspeech 2023*, 2023, pp. 396–400.

[303] L. Zhang, N. Jiang, Q. Wang, Y. Li, Q. Lu, and L. Xie, "Whisper-sv: Adapting whisper for low-data-resource speaker verification," *Speech Communication*, vol. 163, p. 103103, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016763932400075X.

[304] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, Z. Meng, K. Hu, A. Rosenberg, R. Prabhavalkar, D. S. Park, P. Haghani, J. Riesa, G. Perng, H. Soltau, T. Strohman, B. Ramabhadran, T. Sainath, P. Moreno, C.-C. Chiu, J. Schalkwyk, F. Beaufays, and Y. Wu, "Google usm: Scaling automatic speech recognition beyond 100 languages," 2023. [Online]. Available: https://arxiv.org/abs/2303.01037.

[305] I. Tsiamas, D. Dale, and M. R. Costa-jussà, "Improving language and modality transfer in translation by character-level modeling," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 20171–20187. [Online]. Available: https://aclanthology.org/2025.acl-long.988/.

[306] R. Ma, A. Liusie, M. Gales, and K. Knill, "Investigating the emergent audio classification ability of ASR foundation models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 4746–4760. [Online]. Available: https://aclanthology.org/2024.naacl-long.266/.

[307] S. Liang and G.-A. Levow, "Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages," in *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, É. Le Ferrand, E. Klyachko, A. Postnikova, T. Shavrina, O. Serikov, E. Voloshina, and E. Vylomova, Eds. Vienna, Austria: Association for Computational Linguistics, Aug. 2025, pp. 26–37. [Online]. Available: https://aclanthology.org/2025.fieldmatters-1.3/.

[308] I. Kipyatkova, I. Kagirov, M. Dolgushin, and A. Rodionova, "Towards a livvi-karelian end-to-end asr system," in *Speech and Computer*, A. Karpov and V. Delić, Eds. Cham: Springer Nature Switzerland, 2025, pp. 57–68.

[309] D. Mengke, Y. Meng, and P. Mihajlik, "Transliteration-aided transfer learning for low-resource asr: A case study on khalkha mongolian," *Electronics*, vol. 14, no. 6, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/14/6/1137.

[310] L. G. Pillai, K. Manohar, B. K. Raju, and E. Sherly, "Multistage fine-tuning strategies for automatic speech recognition in low-resource languages," 2024. [Online]. Available: https://arxiv.org/abs/2411.04573.

[311] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource ASR," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 29, Jun. 2024.

[312] O. H. Anidjar, R. Marbel, and R. Yozevitch, "Whisper turns stronger: Augmenting wav2vec 2.0 for superior asr in low-resource languages," 2024. [Online]. Available: https://arxiv.org/abs/2501.00425.

[313] R. Coto-Solano, T. W. Kim, A. Jones, and S. Loáiciga, "Multilingual models for ASR in chibchan languages," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 8521–8535. [Online]. Available: https://aclanthology.org/2024.naacl-long.471/.

[314] N. Jakhar, S. Srivastava, and A. Baby, "A Unified Approach to Multilingual Automatic Speech Recognition with Improved Language Identification for Indic Languages," in *Interspeech 2024*, 2024, pp. 3949–3953.

[315] R. Bagat, I. Illina, and E. Vincent, "Mixture of LoRA Experts for Low-Resourced Multi-Accent Automatic Speech Recognition," in *Interspeech 2025*, 2025, pp. 1143–1147.

[316] T. P. Ferraz, "Efficient compression of multitask multilingual speech models," 2024. [Online]. Available: https://arxiv.org/abs/2405.00966.

[317] K. Tripathi, R. Gothi, and P. Wasnik, "Enhancing whispers accuracy and speed for indian languages through prompt-tuning and tokenization," in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[318] J. Li, Y. Pu, Q. Sun, and W.-Q. Zhang, "Improving whispers recognition performance for under-represented language kazakh leveraging unpaired speech and text," in *Interspeech 2024*. ISCA, Sep. 2024, p. 25142518. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2024-1790.

[319] W. Liu, Y. Qin, Z. Peng, and T. Lee, "Sparsely shared lora on whisper for child speech recognition," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11751–11755.

[320] G. I. Gllego, O. Pareras, M. Cortada Garcia, L. Takanori, and J. Hernando, "Speech-to-Text Translation with Phoneme-Augmented CoT: Enhancing Cross-Lingual Transfer in Low-Resource Scenarios," in *Interspeech 2025*, 2025, pp. 31–35.

[321] Y. Yang, Z. Song, J. Zhuo, M. Cui, J. Li, B. Yang, Y. Du, Z. Ma, X. Liu, Z. Wang, K. Li, S. Fan, K. Yu, W.-Q. Zhang, G. Chen, and X. Chen, "GigaSpeech 2: An evolving, large-scale and multi-domain ASR corpus for low-resource languages with automated crawling, transcription and refinement," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J.

Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 2673–2686. [Online]. Available: https://aclanthology.org/2025.acl-long.135/.

[322] Y. Luo and P. Mihajlik, "Comparison of well and lower-resourced self-training in asr," in *Speech and Computer*, A. Karpov and V. Delić, Eds. Cham: Springer Nature Switzerland, 2025, pp. 45–56.

[323] J. H. Yeo, M. Kim, S. Watanabe, and Y. M. Ro, "Visual speech recognition for languages with limited labeled data using automatic labels from whisper," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10471–10475.

[324] C. Chhablani, N. Sharma, J. Hosier, and V. K. Gurbani, "Digits micro-model for accurate and secure transactions," 2024. [Online]. Available: https://arxiv.org/abs/2402.01931.

[325] S. Yang, D. Kim, and S. Lee, "Hylr-fo: Hybrid approach using language models and rule-based systems for on-device food ordering," *Electronics*, vol. 14, no. 4, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/14/4/775.

[326] A. Basit, K. Hussain, M. A. Hanif, and M. Shafique, "Robomed: On-premise medical assistance leveraging large language models in robotics," in *2024 18th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2024, pp. 710–717.

[327] L. Zhang, S. Wu, and Z. Wang, "Lora-int8 whisper: A low-cost cantonese speech recognition framework for edge devices," *Sensors*, vol. 25, no. 17, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/17/5404.

[328] C. Bao, C. Huo, Q. Chen, and C. Gao, "As-asr: A lightweight framework for aphasia-specific automatic speech recognition," 2025. [Online]. Available: https://arxiv.org/abs/2506.06566.

[329] M. Carl and M. Icht, "Automated assessment of word- and sentence-level speech intelligibility in developmental motor speech disorders: A cross-linguistic investigation," *Diagnostics*, vol. 15, no. 15, 2025. [Online]. Available: https://www.mdpi.com/2075-4418/15/15/1892.

[330] D. Rakhimova, Z. Duisenbekkyzy, and E. Adali, "Investigation of asr models for low-resource kazakh child speech: Corpus development, model adaptation, and evaluation," *Applied Sciences*, vol. 15, no. 16, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/16/8989.

[331] R. Jain, A. Barcovschi, M. Y. Yiwere, P. Corcoran, and H. Cucu, "Exploring native and non-native english child speech recognition with whisper," *IEEE Access*, vol. 12, pp. 41601–41610, 2024.

[332] M. Shao, X. Zhu, C. Wang, B. Mu, H. Li, Y. Yan, J. Liu, D. Xie, and L. Xie, "Weakly Supervised Data Refinement and Flexible Sequence Compression for Efficient Thai LLM-based ASR," in *Interspeech 2025*, 2025, pp. 748–752.

[333] A. Mittal, D. Prabhu, S. Sarawagi, and P. Jyothi, "SALSA: Speedy ASR-LLM Synchronous Aggregation," in *Interspeech 2024*, 2024, pp. 3485–3489.

[334] T. Xu, H. Chen, Q. Wang, L. Hang, J. Kang, J. Li, Z. Lin, Y. Li, and L. Xie, "Leveraging LLM and Self-Supervised Training Models for Speech Recognition in Chinese Dialects: A Comparative Analysis," in *Interspeech 2025*, 2025, pp. 584–588.

[335] Z. Song, Z. Ma, Y. Yang, J. Zhuo, and X. Chen, "A comparative study of llm-based asr and whisper in low resource and code switching scenario," 2024. [Online]. Available: https://arxiv.org/abs/2412.00721.

[336] J. Lee and H. Seo, "Online Knowledge Distillation of Decoder-Only Large Language Models for Efficient Speech Recognition," in *Interspeech 2024*, 2024, pp. 2890–2894.

[337] H. Wang, G. Hu, G. Lin, W.-Q. Zhang, and J. Li, "Simul-Whisper: Attention-Guided Streaming Whisper with Truncation Detection," in *Interspeech 2024*, 2024, pp. 4483–4487.

[338] A. Robatian, M. Hajipour, M. R. Peyghan, F. Rajabi, S. Amini, S. Ghaemmaghami, and I. Gholampour, "Gec-rag: Improving generative error correction via retrieval-augmented generation for automatic speech recognition systems," 2025. [Online]. Available: https://arxiv.org/abs/2501.10734.

[339] C. Sicard, V. Gillioz, and K. Pyszkowski, "Spaiche: Extending state-of-the-art ASR models to Swiss German dialects," in *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, H. Ghorbel, M. Sokhn, M. Cieliebak, M. H¨urlimann, E. de Salis, and J. Guerne, Eds. Neuchatel, Switzerland: Association for Computational Linguistics, Jun. 2023, pp. 76–83. [Online]. Available: https://aclanthology.org/2023.swisstext-1.8/.

[340] S. Li, C. Chen, C. Y. Kwok, C. Chu, E. S. Chng, and H. Kawai, "Investigating ASR Error Correction with Large Language Model and Multilingual 1-best Hypotheses," in *Interspeech 2024*, 2024, pp. 1315–1319.

[341] K. D Chaparala, G. Zarrella, B. Torres Fischer, L. Kimura, and O. Parker Jones, "Mai Hoomuna i ka Ai: Language models improve automatic speech recognition in hawaiian," in *Proceedings of The 4th NeurIPS Efficient Natural Language and Speech Processing Workshop*, ser. Proceedings of Machine Learning Research, M. Rezagholizadeh, P. Passban, S. Samiee, V. Partovi Nia, Y. Cheng, Y. Deng, Q. Liu, and B. Chen, Eds., vol. 262. PMLR, 14 Dec 2024, pp. 576–583. [Online]. Available: https://proceedings.mlr.press/v262/d-chaparala24a.html.

[342] X. de Zuazo, E. Navas, I. Saratxaga, and I. H. Rioja, "Whisper-lm: Improving asr models with language models for low-resource languages," 2025. [Online]. Available: https://arxiv.org/abs/2503.23542.

[343] J. Zhou, S. Zhao, J. He, H. Wang, W. Zeng, Y. Chen, H. Sun, A. Kong, and Y. Qin, "M2r-whisper: Multi-stage and multi-scale retrieval augmentation for enhancing whisper," in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[344] A. A. Abdullah, S. Tabibian, H. Veisi, A. Mahmudi, and T. Rashid, "End-to-end transformer-based automatic speech recognition for northern kurdish: A pioneering approach," 2024. [Online]. Available: https://arxiv.org/abs/2410.16330.

[345] S. Saranya, B. Bharathi, S. Gomathy Dhanya, and A. Krishnakumar, "Real-Time continuous tamil dialect speech recognition and summarization," *Circuits, Systems, and Signal Processing*, vol. 44, no. 4, pp. 2855–2881, Apr. 2025.

[346] mer Tarik zyilmaz, M. Coler, and M. Valdenegro-Toro, "Overcoming Data Scarcity in Multi-Dialectal Arabic ASR via Whisper Fine-Tuning," in *Interspeech 2025*, 2025, pp. 1158–1162.

[347] H. Polat, A. K. Turan, C. Koak, and H. B. UlaÅ, "Implementation of a whisper architecture-based turkish automatic speech recognition (asr) system and evaluation of the effect of fine-tuning with a low-rank adaptation (lora) adapter on its performance," *Electronics*, vol. 13, no. 21, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/21/4227.

[348] D. K. Gete, B. Y. Ahmed, T. D. Belay, Y. A. Ejigu, S. H. Imam, A. B. Tessema, M. O. Adem, T. A. Belay, R. Geislinger, U. A. Musa, M. Semmann, S. H. Muhammad, H. Schreiber, and S. M. Yimam, "Whispering in amharic: Fine-tuning whisper for low-resource language," 2025. [Online]. Available: https://arxiv.org/abs/2503.18485.

[349] S. Arif, A. J. Khan, M. Abbas, A. A. Raza, and A. Athar, "WER we stand: Benchmarking Urdu ASR models," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 5952–5961. [Online]. Available: https://aclanthology.org/2025.coling-main.397/.

[350] X. Li, G. Zhang, A. Zhu, W. Li, S. Fang, X. Yang, and J. Zhu, "Transsion tsup's speech recognition system for asru 2023 madasr challenge," 2023. [Online]. Available: https://arxiv.org/abs/2307.11778.

[351] N.-D. Mai, M.-T. Nguyen, V.-H. Nguyen, X.-B. Le, H.-T. Nguyen, K. Satoh, H. Takeda, and T.-H.-Y. Vuong, "Towards accurate legal term detection: Insights fromdependency tree-based andlarge language model approaches," in *New Frontiers in Artificial Intelligence*, Y. Nakano and T. Suzumura, Eds. Singapore: Springer Nature Singapore, 2025, pp. 3–19.

[352] H. Futami, J. Huynh, S. Arora, S.-L. Wu, Y. Kashiwagi, Y. Peng, B. Yan, E. Tsunoo, and S. Watanabe, "The pipeline system of asr and nlu with mlm-based data augmentation toward stop low-resource challenge," in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–2.

[353] Y. Perezhohin, T. Santos, V. Costa, F. Peres, and M. Castelli, "Enhancing automatic speech recognition: Effects of semantic audio filtering on models performance," *IEEE Access*, vol. 12, pp. 155136–155150, 2024.

[354] M. Qian, S. Tang, R. Ma, K. M. Knill, and M. J. Gales, "Learn and Don't Forget: Adding a New Language to ASR Foundation Models," in *Interspeech 2024*, 2024, pp. 2544–2548.

[355] A. Prasad, S. Madikeri, D. Khalil, P. Motlicek, and C. Schuepbach, "Speech and Language Recognition with Low-rank Adaptation of Pretrained Models," in *Interspeech 2024*, 2024, pp. 2825–2829.

[356] Q. Zhao, G. Sun, C. Zhang, M. Xu, and T. F. Zheng, "Enhancing quantised end-to-end asr models via personalisation," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12426–12430.

[357] Q. Zhao, G. Sun, C. Zhang, M. Xu and T. F. Zheng, "Speaker Adaptive Mixture of Weight-Decomposed LoRA Experts for On-Device End-to-End ASR," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2485–2496, 2025.

[358] G. Ramesh, K. Audhkhasi, and B. Ramabhadran, "Task vector algebra for asr models," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12256–12260.

[359] N. B. Shankar, Z. Wang, E. Eren, and A. Alwan, "Selective attention merging for low resource tasks: A case study of child asr," in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[360] S. Zaiem, T. Parcollet, and S. Essid, "Automatic data augmentation for domain adapted fine-tuning of self-supervised speech representations," in *Interspeech 2023*, 2023, pp. 67–71.

[361] M. U. Farooq and T. Hain, "Learning cross-lingual mappings for data augmentation to improve low-resource speech recognition," in *Interspeech 2023*, 2023, pp. 5072–5076.

[362] W. Wang and Y. Qian, "Unisplice: Universal cross-lingual data splicing for low-resource asr," in *Interspeech 2023*, 2023, pp. 2253–2257.

[363] W. Wang and Y. Qian, "Universal Cross-Lingual Data Generation for Low Resource ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 973–983, 2024.

[364] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Data augmentation for asr using tts via a discrete representation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 68–75.

[365] Y. Matsuzaka, R. Takashima, C. Sasaki, and T. Takiguchi, "Data augmentation for dysarthric speech recognition based on text-to-speech synthesis," in *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, 2022, pp. 399–400.

[366] G. Yang, F. Yu, Z. Ma, Z. Du, Z. Gao, S. Zhang, and X. Chen, "Enhancing low-resource asr through versatile tts: Bridging the data gap," in *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[367] G. Karakasidis, N. Robinson, Y. Getman, A. Ogayo, R. Al-Ghezi, A. Ayasi, S. Watanabe, D. R. Mortensen, and M. Kurimo, "Multilingual tts accent impressions for accented asr," in *Text, Speech, and Dialogue*, K. Ekštein, F. Pártl, and M. Konopík, Eds. Cham: Springer Nature Switzerland, 2023, pp. 317–327.

[368] Z. Huang, G. Keren, Z. Jiang, S. Jain, D. Goss-Grubbs, N. Cheng, F. Abtahi, D. Le, D. Zhang, A. D'Avirro, E. Campbell-Taylor, J. Salas, I.-E. Veliche, and X. Chen, "Text generation with speech synthesis for asr data augmentation," 2023. [Online]. Available: https://arxiv.org/abs/2305.16333.

[369] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin, "You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2020, pp. 439–444.

[370] Z. Jin, X. Xie, M. Geng, T. Wang, S. Hu, J. Deng, G. Li, and X. Liu, "Adversarial data augmentation using vae-gan for disordered speech recognition," in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[371] S. Huang, D. Zhang, J. Deng, and R. Zheng, "Enhanced asr for stuttering speech: Combining adversarial and signal-based data augmentation," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 393–400.

[372] P. Sheng, Z. Yang, and Y. Qian, "Gans for children: A generative data augmentation strategy for children speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 129–135.

[373] P. Bawa, V. Kadyan, and G. Chhabra, "Generative adversarial network-based voice synthesis from spectrograms for low-resource speech recognition in mismatched conditions," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–6.

[374] Z. Chen, A. Rosenberg, Y. Zhang, G. Wang, B. Ramabhadran, and P. J. Moreno, "Improving speech recognition using gan-based speech synthesis and contrastive unspoken text selection," in *Interspeech 2020*, 2020, pp. 556–560.

[375] M. Giollo, D. Gunceler, Y. Liu, and D. Willett, "Bootstrap an end-to-end asr system by multilingual training, transfer learning, text-to-text mapping and synthetic audio," in *Interspeech 2021*, 2021, pp. 2416–2420.

[376] L. Jiang, X. Qin, J. Zhang, and J. Li, "Multimodal seed data augmentation for low-resource audio latin cuengh language," *Applied Sciences*, vol. 14, no. 20, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/20/9533.

[377] G. Sidiropoulos and E. Kanoulas, "A multimodal dense retrieval approach for speech-based open-domain question answering," 2024. [Online]. Available: https://arxiv.org/abs/2409.13483.

[378] V. R. Konda, M. Warialani, R. P. Achari, V. Bhatnagar, J. Akula, P. Jyothi, G. Ramakrishnan, G. Haffari, and P. Singh, "Caption alignment for low resource audio-visual data," in *Interspeech 2020*, 2020, pp. 3525–3529.

[379] D. Porjazovski, T. Grsz, and M. Kurimo, "Topic identification for spontaneous speech: Enriching audio features with embedded linguistic information," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 396–400.

[380] P. Gambhir, A. Dev, P. Bansal, and D. K. Sharma, "End-to-end multi-modal low-resourced speech keywords recognition using sequential conv2d nets," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 23, no. 1, Jan. 2024. [Online]. Available: https://doi.org/10.1145/3606019.

[381] S. Kalantari, D. Dean, and S. Sridharan, "Cross database audio visual speech adaptation for phonetic spoken term detection," *Computer Speech & Language*, vol. 44, pp. 1–21, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230815300139.

[382] A. Haliassos, P. Ma, R. Mira, S. Petridis, and M. Pantic, "Jointly learning visual and auditory speech representations from raw data," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=BPwIgvf5iQ.

[383] W. Dai, S. Cahyawijaya, T. Yu, E. J. Barezi, P. Xu, C. T. Yiu, R. Frieske, H. Lovenia, G. Winata, Q. Chen, X. Ma, B. Shi, and P. Fung, "CI-AVSR: A Cantonese audio-visual speech datasetfor in-car command recognition," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6786–6793. [Online]. Available: https://aclanthology.org/2022.lrec-1.731/.

[384] C. Yu, J. Yu, Z. Qian, and Y. Tan, "Improvement of acoustic models fused with lip visual information for low-resource speech," *Sensors*, vol. 23, no. 4, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/4/2071.

[385] D. Dash, M. Kim, K. Teplansky, and J. Wang, "Automatic speech recognition with articulatory information and a unified dictionary for hindi, marathi, bengali and oriya," in *Interspeech 2018*, 2018, pp. 1046–1050.

[386] J. Peymanfard, S. Heydarian, A. Lashini, H. Zeinali, M. R. Mohammadi, and N. Mozayani, "A multi-purpose audio-visual corpus for multi-modal persian speech recognition: The arman-av dataset," *Expert Systems with Applications*, vol. 238, p. 121648, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423021504.

[387] A. Gupta, A. Parulekar, S. Chattopadhyay, and P. Jyothi, "Parameter-efficient adaptation of multilingual multimodal models for low-resource ASR," in *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, J. Sälevä and A. Owodunni, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 175–185. [Online]. Available: https://aclanthology.org/2024.mrl-1.13/.

[388] S. Mdhaffar, H. Elleuch, F. Bougares, and Y. Estève, "Performance analysis of speech encoders for low-resource SLU and ASR in Tunisian dialect," in *Proceedings of the Second Arabic Natural Language Processing Conference*, N. Habash, H. Bouamor, R. Eskander, N. Tomeh, I. Abu Farha, A. Abdelali, S. Touileb, I. Hamed, Y. Onaizan, B. Alhafni, W. Antoun, S. Khalifa, H. Haddad, I. Zitouni, B. AlKhamissi, R. Almatham, and K. Mrini, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 130–139. [Online]. Available: https://aclanthology.org/2024.arabicnlp-1.12/.

[389] C. Meng and A. Anastasopoulos, "GMU systems for the IWSLT 2025 low-resource speech translation shared task," in *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, E. Salesky, M. Federico, and A. Anastasopoulos, Eds. Vienna, Austria (in-person and online): Association for Computational Linguistics, Jul. 2025, pp. 289–300. [Online]. Available: https://aclanthology.org/2025.iwslt-1.29/.

[390] Y. Du, Y. Pan, Z. Ma, B. Yang, Y. Yang, K. Deng, X. Chen, Y. Xiang, M. Liu, and B. Qin, "Making llms better many-to-many speech-to-text translators with curriculum learning," 2025. [Online]. Available: https://arxiv.org/abs/2409.19510.

[391] X. Wang, Y. Li, C. Fu, Y. Zhang, Y. Shen, L. Xie, K. Li, X. Sun, and L. MA, "Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=s1EImzs5Id.

[392] J. Chen, C. Liu, and M. Li, "Automatic emotional spoken language text corpus construction from written dialogs in fictions," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 319–324.

[393] Z. Yang, C. Lydia, M. Kathleen, M. Smaranda, and R. Andrew, "Identifying speaker state from multimodal cues," Ph.D. dissertation, Columbia University, USA, 2021, aAI28770950.

[394] S. Ouali and S. E. Garouani, "Mder-ma: A multimodal dataset for emotion recognition in low-resource moroccan arabic language," *Data in Brief*, vol. 62, p. 112005, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340925007292.

[395] AhmedTanveer, M. Ali, HussainSaqib, TanwaniAmer, and SadatAhmed, "Emotion recognition of low resource (sindhi) language using machine learning," *International Journal of Computer Science and Network Security*, vol. 21, no. 8, pp. 369–376, 82021.

[396] W. Dai, S. Cahyawijaya, T. Yu, E. J. Barezi, and P. Fung, "Kaggle competition: Cantonese audio-visual speech recognition for in-car commands," 2022. [Online]. Available: https://arxiv.org/abs/2207.02663.

[397] M. Marreddy, S. R. Oota, L. S. Vakada, V. C. Chinni, and R. Mamidi, "Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 1, Nov. 2022. [Online]. Available: https://doi.org/10.1145/3531535.

[398] C. Wang, A. Wu, J. Gu, and J. Pino, "Covost 2 and massively multilingual speech translation," in *Interspeech 2021*, 2021, pp. 2247–2251.

[399] 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.

[400] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "The multilingual tedx corpus for speech recognition and translation," in *Interspeech 2021*, 2021, pp. 3655–3659.

[401] P. P. S. Aaron M. Oirere, Ratnadeep R. Deshmukh, "Development of isolated numeric speech corpus for swahili language for development of automatic speech recognition system," *International Journal of Computer Applications*, vol. 74, no. 11, pp. 20–22, July 2013. [Online]. Available: https://ijcaonline.org/archives/volume74/number11/12929-9841/.

[402] K. Soky, Z. Gong, and S. Li, "Nict-tib1: A public speech corpus of lhasa dialect for benchmarking tibetan language speech recognition systems," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2022, pp. 1–5.

[403] U. Kimanuka, C. wa Maina, and O. Byk, "Speech recognition datasets for low-resource congolese languages," *Data in Brief*, vol. 52, p. 109796, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340923008582.

[404] A. Mukhamadiyev, M. Mukhiddinov, I. Khujayarov, M. Ochilov, and J. Cho, "Development of language models for continuous uzbek speech recognition system," *Sensors*, vol. 23, no. 3, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/3/1145.

[405] X. Li, D. R. Mortensen, F. Metze, and A. W. Black, "Multilingual phonetic dataset for low resource speech recognition," in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6958–6962.

## AUTHOR BIOGRAPHY



**Donghong Qin** received his Ph.D. degree in Computer Science and Technology from Tsinghua University, China. He was a visiting scholar at the University of Massachusetts, U.S., and a visiting professor at Tsinghua University. His research focuses on natural language processing, machine translation, big data processing, and intelligent algorithms. He has led several national and provincial/ministerial projects and published over 80 papers, including more than 40 SCI/EIindexed articles.
ORCID: 0000-0003-1448-2162



**Zhuo Chen** received his bachelor's degree from Henan Normal University, China. He is currently pursuing the master's degree at Guangxi Minzu University, China, and is a Student Member of the China Computer Federation (CCF). His research interests include natural language processing, low-resource automatic speech recognition, and multimodal speech processing, with emphasis on speech representation learning, efficient audio-visual feature integration, and automated neural architecture design for speech and language models.
ORCID: 0009-0007-9275-6436



**Xiao Liang** received his B.S. degree in Electronic Information Engineering from Xidian University, China, his M.Sc. degree in Electrical Engineering and Information Technology from Leibniz University Hannover, Germany, and his Ph.D. degree in Computer Science from Universiti Tunku Abdul Rahman, Malaysia. His research interests include artificial intelligence, natural language processing, machine translation, multilingual speech translation, and parameter-efficient fine-tuning of large language models. He has published several papers in domestic and international journals, including two SCI-indexed articles.
ORCID: 0009-0006-5321-4565



**Fengbo Bai** received his M.Sc. degree in Artificial Intelligence from the University of St Andrews, U.K., and his interdisciplinary Ph.D. degree in Evidence Law from China University of Political Science and Law, China. His research focuses on computational linguistics and artificial intelligence. He has led or participated in several national and provincial/ministerial research projects and published nearly 30 papers in major journals and conferences, including 16 indexed by core journals, SCI, or EI. He has also filed five invention patents.
ORCID: 0009-0009-1534-6868