IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 April 2025, accepted 18 June 2025, date of publication 23 June 2025, date of current version 1 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3582395

SURVEY

A Survey on Data Selection for Efficient Speech Processing

ABDUL HAMEED AZEEMI, IHSAN AYYUB QAZI, AND AGHA ALI RAZA
Department of Computer Science, Lahore University of Management Sciences, Lahore 54792, Pakistan
Corresponding author: Abdul Hameed Azeemi (e-mail: abdul.azeemi@lums.edu.pk)

This work was supported by LUMS.

ABSTRACT While recent advances in speech processing have led to substantial performance improvements across diverse tasks, they often demand significantly higher computational costs and resources. To address this efficiency challenge, data selection has emerged as a crucial strategy. This survey provides a comprehensive overview and introduces a unifying taxonomy for data selection methods in speech processing, structured along three key dimensions: selection granularity (sample-level vs. segment-level), selection process (static, dynamic, or active learning), and selection criteria (uncertainty, diversity, or hybrid approaches). Through systematic analysis across major speech tasks, including automatic speech recognition, text-to-speech synthesis, audio anti-spoofing, speaker recognition, and emotion recognition, we evaluate the effectiveness and applicability of diverse data selection strategies. Our analysis reveals that targeted data selection not only alleviates computational burdens but often enhances model robustness and performance by strategically filtering redundant, noisy, or detrimental training examples. By synthesizing insights scattered across disparate speech domains, we identify common principles, highlight task-specific challenges, and reveal emerging research trends. Finally, we outline promising future research directions in data selection for efficient speech processing.

INDEX TERMS Speech processing, survey, data selection, data pruning, active learning, computational efficiency, data-efficient learning.

I. INTRODUCTION

Speech processing technologies have evolved substantially over the past few decades, transitioning from early rule-based systems with constrained vocabularies to sophisticated end-to-end neural architectures [1]. State-of-the-art speech models, such as Whisper [2] for automatic speech recognition and VALL-E 2 [3] for text-to-speech, have set new benchmarks, demonstrating robustness and effectiveness across diverse languages and challenging acoustic conditions.

However, these gains in performance have not come without substantial computational costs. As speech models grow in scale and complexity, their training increasingly demands massive computational resources and extensive datasets. For instance, the Whisper (large) model required training on approximately 680,000 hours of transcribed speech [2], while VALL-E 2 relied on 50,000 hours of English speech data [3]. This trend is consistent with neural scaling laws, which indicate a power-law relationship between model performance, dataset size, and computational resources [4]. Consequently, these large-scale training regimens pose several critical challenges: prohibitive financial costs that restrict state-of-the-art model development to well-funded institutions, prolonged training cycles that impede rapid experimentation and innovation, and significant environmental consequences due to substantial energy consumption and associated carbon emissions [5]. As speech systems continue to scale, these factors may widen the divide between resource-rich and resource-constrained environments, potentially limiting innovation and diversity in the field.

To mitigate these computational demands, data selection has emerged as a promising strategy for maintaining model performance while substantially improving training efficiency. Interest in data selection methodologies across various speech processing tasks has steadily increased, driven by the need for more computationally efficient training

The associate editor coordinating the review of this manuscript and approving it for publication was Tony Thomas.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

paradigms. At its core, data selection addresses a fundamental challenge in machine learning: How can we effectively construct smaller yet informative datasets from large, uncurated collections of raw data? [6]. Although traditional wisdom in speech processing has favored the notion that *"more data is better,"* recent studies demonstrate that selectively removing redundant, noisy, or less informative examples can not only reduce computational overhead but also maintain, and occasionally even improve, overall model performance.

A. RESEARCH GAP

Despite the growing body of work on data selection across various speech processing tasks, there is a lack of a unifying framework that systematically categorizes and analyzes these approaches. Existing literature on data selection is fragmented across different domains of speech processing, making it difficult to identify common principles, best practices, and transferable methodologies. Additionally, while many studies report efficiency gains from data selection, there is no comprehensive comparison of different selection strategies across comparable benchmarks and evaluation metrics.

B. MOTIVATION AND OBJECTIVES

Our primary motivation for conducting this survey is to develop a comprehensive taxonomy of methods that address the escalating computational and resource challenges in speech processing. As models continue to grow in scale, the environmental impact, financial cost, and accessibility issues associated with large-scale training become increasingly significant concerns. Effective data selection offers a pathway to more sustainable and inclusive speech technology development by reducing training data requirements without sacrificing performance. Furthermore, understanding which data selection techniques are most effective across different speech tasks can accelerate research progress in different domains of speech processing.

C. CONTRIBUTIONS AND ORGANIZATION

In this survey, we:

- Develop and present a novel, unifying taxonomy for data selection methods in speech processing, categorizing them based on granularity, process, and criteria to provide a structured understanding of the field.
- Provide a comprehensive review of data selection techniques across key speech processing tasks: automatic speech recognition, text-to-speech synthesis, audio anti-spoofing, speaker recognition, and emotion recognition.
- Systematically analyze and synthesize the limitations, challenges, and open research questions associated with data selection within each speech domain, identifying commonalities and task-specific nuances.
- Identify and outline promising future research directions, highlighting underexplored areas in data selection for efficient and effective speech processing.

The rest of the paper is organized as follows. Section IV presents a taxonomy of data pruning methods. We then review data selection for specific speech processing tasks: Section V-A covers automatic speech recognition, Section V-B explores text-to-speech systems, Section V-D examines audio anti-spoofing, Section V-E discusses speaker recognition, Section V-F analyzes emotion recognition, and Section V-G outlines data selection for other speech tasks (Keyword Spotting and Speaker Diarization). Finally, Section VI concludes the survey with a summary of findings and broader implications for the field.

II. BACKGROUND AND PRELIMINARIES

We now establish the definitions and concepts related to data selection in speech processing. These concepts provide a framework for understanding how data selection operates on speech data and form the basis for the taxonomy of data selection methods presented in Section IV.

A. DEFINITIONS

1) SPEECH DATA POINT

A speech data point, $x^{(i)}$, is a discrete unit of speech data used to train or evaluate a speech processing model. Depending on the task and method, this may be a single utterance or a speech segment. For automatic speech recognition, a data point typically consists of an audio recording paired with its transcription. For speaker recognition, it might be an utterance with a speaker identity label. The granularity of a data point varies based on the specific task requirements.

2) SPEECH DATASET

A speech dataset, D, is a collection of speech data points $\{x^{(1)}, \ldots, x^{(N)}\}$ (where $N = |D|$ is the size of the dataset) used to train or evaluate a speech processing model. Speech datasets often require substantial resources to collect and annotate, especially for supervised learning tasks that need accurate transcriptions or labels. This resource intensity makes efficient data selection particularly valuable in speech processing domains.

3) DATA UTILITY FUNCTION

A data utility function, $U(x^{(i)})$, assigns a value to each speech data point $x^{(i)}$ that quantifies its usefulness for training a particular model for a specific task. The utility function may measure various attributes, including the information content of the data point, its representativeness of the target distribution, its uniqueness within the dataset, or its expected contribution to model performance improvement. Different data selection methods effectively implement different utility functions.

4) DATA SELECTION

Data selection is the process of identifying a subset $S \subset D$ of the original dataset that maximizes some objective function while satisfying constraints on the subset size $|S| \leq k$,

where k is typically much smaller than $|D|$. The objective function often involves maximizing the cumulative utility of the selected data points while ensuring diversity and coverage of the relevant data distribution. Formally, this can be expressed as:

$$S^* = \underset{S \subset D, |S| \leq k}{\operatorname{argmax}} \sum_{x^{(i)} \in S} U(x^{(i)}) \tag{1}$$

In practice, finding the optimal subset S^* is computationally intractable for large datasets, leading to the development of various approximation algorithms and heuristics that we explore in this survey.

In this work, we include data pruning (the process of removing less valuable samples from a dataset) and active learning (the strategic selection of unlabeled samples for annotation) as methodologies within the broader data selection paradigm, as both of these involve making decisions about data utility and selecting optimal subsets of available data.

Data selection shares conceptual similarities with feature selection in classical machine learning [7], as both aim to enhance efficiency. However, while feature selection seeks to identify the most informative *input features*, data selection usually focuses on identifying the most informative *data points*.

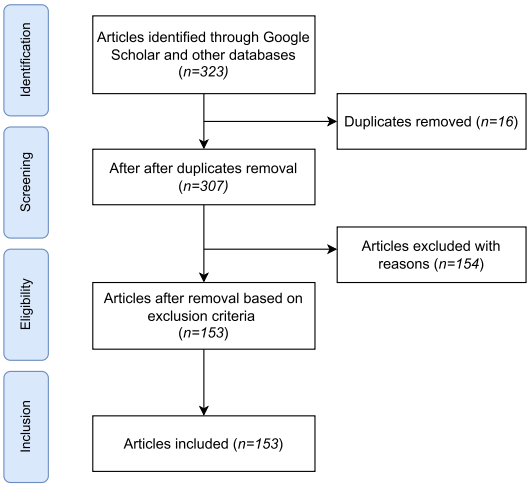

FIGURE 1. Chart illustrating the paper selection process for this survey according to PRISMA guidelines.

III. METHODOLOGY

In this section, we describe our approach for selecting papers for this survey, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [8]. The PRISMA framework ensures transparency, reproducibility, and methodological rigor in our literature search and paper selection process.

A. SEARCH STRATEGY

We conducted a comprehensive search of the scientific literature on data selection methods across various speech processing tasks. The search was performed using different academic databases, including Google Scholar, Dimensions.ai, IEEE Xplore, ACM Digital Library, Springer Link, ScienceDirect, and DBLP. These databases were chosen to ensure broad coverage of relevant research in speech technology, machine learning, and signal processing. Given the diverse terminology used across different speech processing domains, we constructed task-specific search queries combining general data selection terms with speech task descriptors. This approach allowed us to capture relevant studies across the spectrum of speech processing tasks while maintaining specificity to data selection methodologies.

B. SEARCH QUERIES

We formulated targeted search queries for specific speech processing tasks examined in this survey. Each query combined data selection methodology terms with task-specific terminology. We also used a broader search query at the end to include selection methods not falling under prior queries.

- **Automatic Speech Recognition (ASR):** (“data selection” OR “active learning” OR “data pruning” OR “subset selection”) AND (“automatic speech recognition” OR “ASR”)
- **Text-to-Speech Synthesis (TTS):** (“data selection” OR “active learning” OR “data pruning” OR “subset selection”) AND (“text to speech” OR “TTS” OR “text-to-speech”)
- **Audio Anti-Spoofing:** (“data selection” OR “active learning” OR “data pruning” OR “subset selection”) AND (“spoof” OR “anti-spoofing” OR “spoofing” OR “antispoofing”)
- **Speaker Recognition:** (“data selection” OR “active learning” OR “data pruning” OR “subset selection”) AND (“speaker recognition”)
- **Emotion Recognition:** (“data selection” OR “active learning” OR “data pruning” OR “subset selection”) AND (“emotion recognition”)
- **Speech Processing:** (“data selection” OR “active learning” OR “data pruning” OR “subset selection”) AND (“speech processing”)

C. EXCLUSION CRITERIA

To ensure the relevance and quality of the selected papers, we established specific exclusion criteria. We excluded papers that focused solely on model architecture, training algorithm improvements, model pruning, or other efficiency techniques that *do not* include data selection, data pruning, or active learning. Additionally, we did not consider publications that were not available in English.

D. SELECTION PROCESS

The selection process followed the PRISMA guidelines and consisted of several steps, as illustrated in Figure 1:

1) **Identification:** We executed the search queries across multiple databases, showing a total of 323 potentially relevant papers.

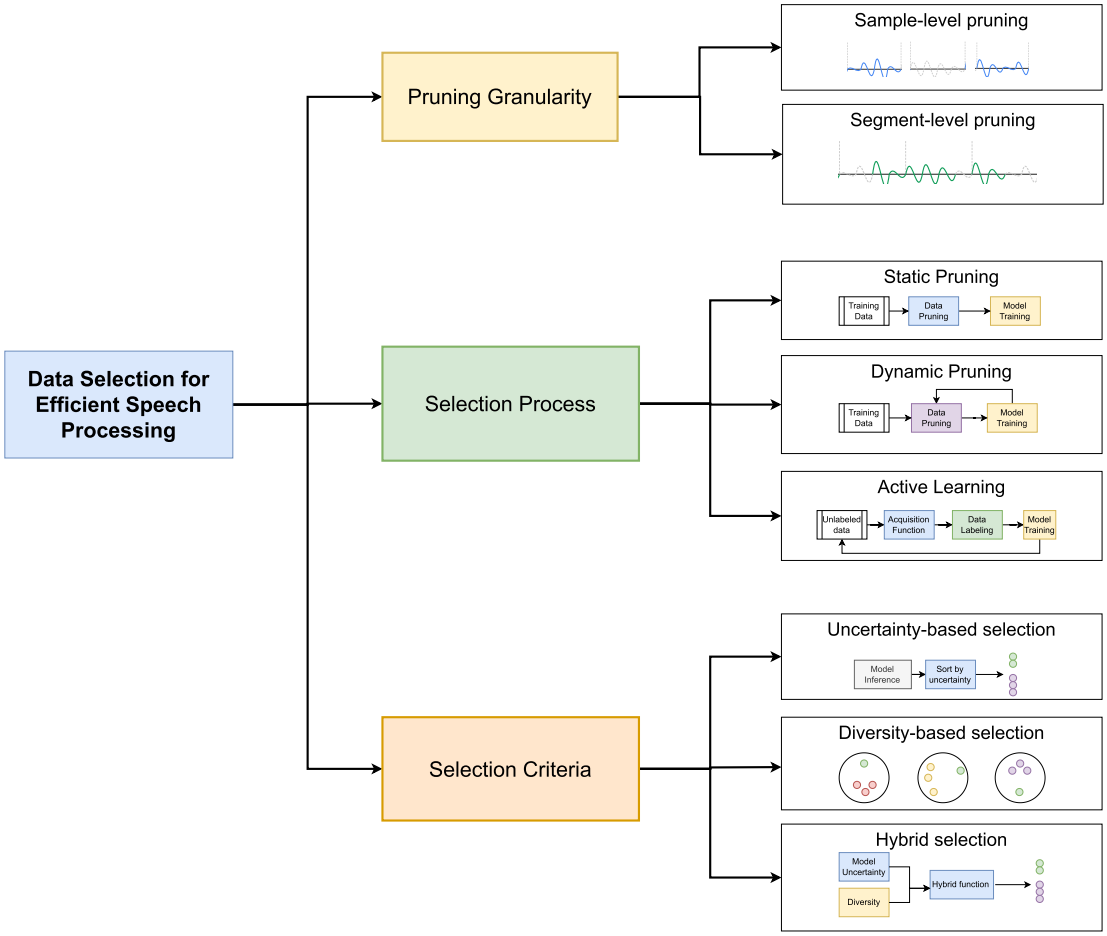

FIGURE 2. Taxonomy of data selection algorithms for speech processing.

2) **Screening:** We identified and removed 16 duplicate entries, resulting in 307 unique records for further review.
3) **Eligibility:** We screened titles and abstracts against the inclusion and exclusion criteria, excluding 154 papers that did not meet our requirements.
4) **Inclusion:** The remaining 153 articles were retained for full-text review and inclusion in our systematic analysis.

The final set of 153 papers forms the core literature base for this survey, representing the state-of-the-art in data selection methodologies across different speech processing tasks. After completing the selection process, we identified common themes and categorized the papers according to different taxonomy dimensions (pruning granularity, selection process, and selection criteria), which we discuss in the next section. This categorization enabled us to analyze trends, identify common principles, and highlight task-specific challenges and opportunities.

IV. TAXONOMY OF DATA PRUNING

Our taxonomy of data pruning methods in speech processing (Fig. 2) emerges from a systematic analysis of the literature, revealing three key dimensions that characterize different approaches: pruning granularity (how much data is considered at once), selection process (when and how data is chosen), and selection criteria (what metrics guide the choice). This categorization reflects the practical considerations that drive different methodological choices. By organizing methods along these dimensions, we can better understand the trade-offs between different approaches.

A. PRUNING GRANULARITY

- **Sample-Level Pruning:** Operates on complete utterances or samples (typically sentences or phrases bounded by silence) to determine which ones to include in the dataset. This approach focuses on selecting representative samples that provide good coverage of speakers, accents, and acoustic conditions. For instance, cluster-based pruning using k-means clustering has been applied to utterance-level audio data, significantly reducing dataset size while maintaining classification performance [9].
- **Segment-Level Pruning:** Works with meaningful contiguous sequences within utterances, such as phonetic

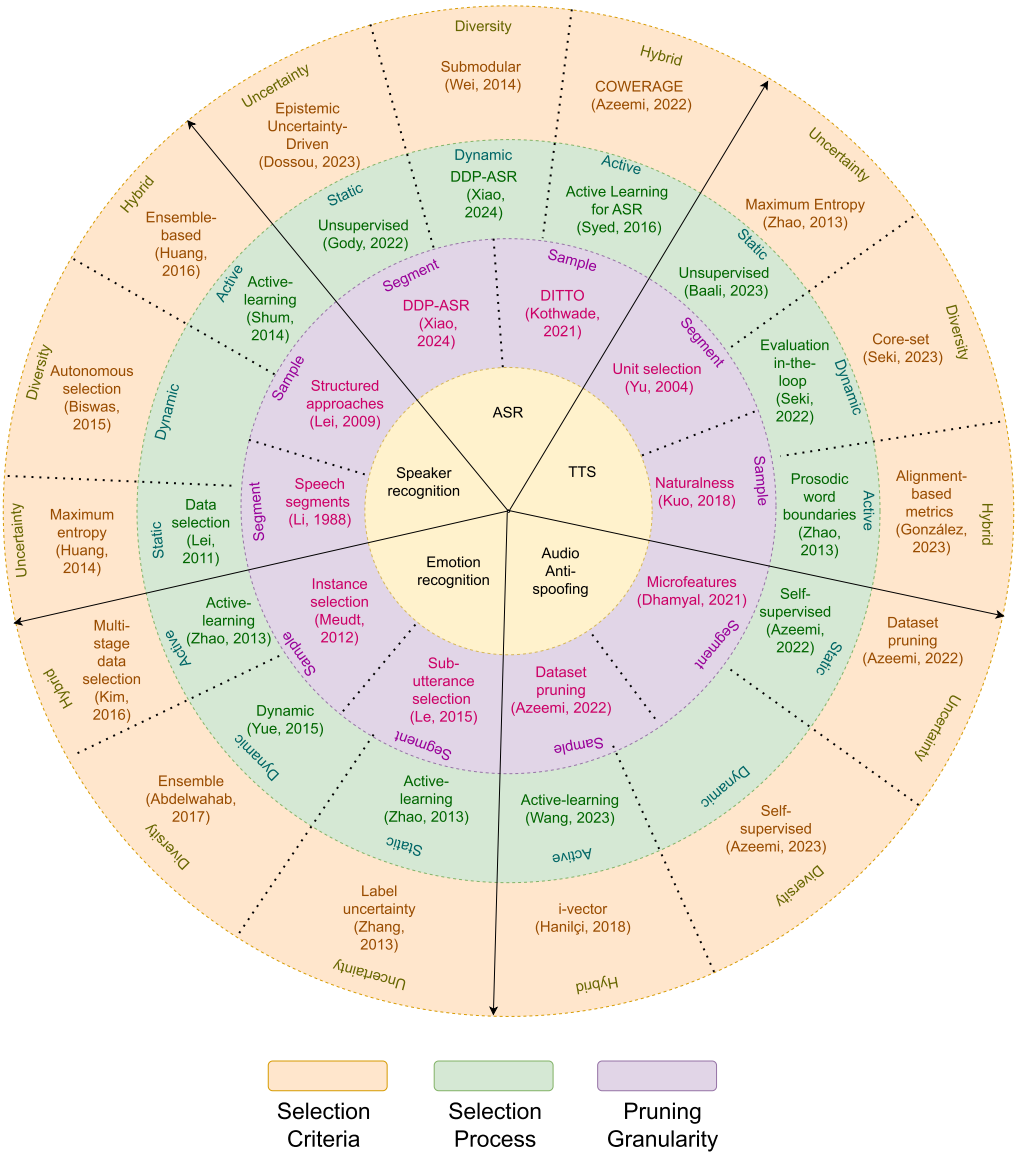

FIGURE 3. A radial map showing examples of the data selection strategies for each task (automatic speech recognition, text-to-speech, audio anti-spoofing, emotion recognition, and speaker recognition) and category (selection granularity, selection process, and selection criteria).

units, words, tokens, or other linguistic elements. This finer-grained approach allows for selective pruning of specific parts within an utterance - for example, keeping emotionally salient segments for emotion recognition while discarding neutral segments. SpeechPrune [10] demonstrated the effectiveness of this approach, achieving significant accuracy improvements at high pruning rates by identifying and discarding less relevant tokens. While segment-level studies are more limited than sample-level ones, this granular approach is particularly valuable for tasks requiring precise focus on specific acoustic features within longer sequences.

B. SELECTION PROCESS

- **Static Pruning:** Selects a subset of data *once* before training based on predefined scores or metrics, such as error rates, uncertainty estimates, or importance scores. By reducing the dataset size upfront, static pruning can significantly decrease training time and computational costs, making it particularly useful in resource-constrained environments. However, its effectiveness heavily depends on the choice of the pruning metric, as an inappropriate selection might lead to suboptimal model performance. While static pruning is widely used for efficiency, it does not adapt

dynamically to the evolving learning needs of the model.

- **Dynamic Pruning:** Involves selecting data on the fly during training based on current model performance, offering flexibility and adaptability. By continuously adjusting the dataset composition, dynamic pruning can focus training on the most informative or challenging examples, improving learning efficiency. The utility function that quantifies the usefulness of a particular data point is usually time-varying for dynamic pruning, i.e., $U(x^{(i)}, t)$, where the selection inherently depends on the model state at time t. This method has been successfully applied in different tasks, such as automatic speech recognition (ASR), where dynamic data pruning can save up to 1.6x training time [11]. However, its effectiveness depends on well-designed selection criteria, as frequent updates can introduce computational overhead.
- **Active Learning:** Active learning frameworks iteratively select the most informative unlabeled examples to annotate, thereby constructing an efficient training subset. Typically, selection is based on model uncertainty or expected impact on the model. For example, in speech tasks, active learning has been used to choose which utterances to transcribe from a large pool, showing better accuracy with less labeled data [12]. The model's current confidence or entropy on each candidate sample guides the selection so that each newly added sample maximally reduces uncertainty. Active learning approaches can dramatically reduce the amount of speech data that needs manual labeling while maintaining recognition performance.

C. SELECTION CRITERIA

- **Uncertainty and Entropy-Based Selection:** Uncertainty measures, such as prediction entropy or confidence scores, guide the pruning or selection of data by identifying the most and least informative samples. High-entropy samples (where the model is unsure) are often the most informative for training, whereas low-entropy (very easy or redundant) samples can be pruned. This principle overlaps with active learning but can also apply in a fully supervised setting – e.g., discarding utterances that a preliminary model already handles with high confidence to focus training on harder cases. The utility function $U(x^{(i)})$ here is often instantiated as the model's uncertainty or entropy for sample $x^{(i)}$. In practice, entropy-based selection has been applied to ensure broad coverage of content. For instance, in text-to-speech corpus design, selecting sentences that maximize the entropy of phonetic/prosodic contexts leads to a more balanced and informative dataset [13]. By prioritizing diverse or information-rich examples, entropy-based pruning avoids over-representing repetitive patterns.
- **Diversity and Coverage-Based Methods**: One major category of subset selection focuses on maximizing coverage of the data distribution. Techniques like submodular selection, clustering, or distribution matching fall into this class. The idea is to pick a subset that is as representative as possible of the whole data – covering various speakers, phonetic units, acoustic conditions, etc. Here, the utility $U(x^{(i)})$ is often defined by the overall objective function being optimized (e.g., maximizing a submodular function). The contribution of a sample depends on the context of other selected samples. Submodular function optimization has been used to select ASR training subsets [14] with near-optimal coverage of the acoustic space. Similarly, ensuring the selected subset matches the target distribution [15] of speech units (phonemes, words, etc.) can show a compact dataset without performance loss. These methods often use distance-based or information coverage criteria – for example, choosing examples that maximize phoneme and prosody coverage in TTS corpora or selecting diverse speakers for multi-speaker models. Diversity-based selection excels at reducing redundancy: it tends to prune duplicates or highly similar samples while keeping outliers that increase variety. The result is an efficient set that maintains robustness across different speaking styles, languages, or conditions.
- **Hybrid Selection:** Hybrid selection methods integrate multiple criteria, combining uncertainty or entropy measures with diversity, coverage, and even adversarial or hard-example considerations to leverage the strengths of each approach. For instance, a hybrid strategy may first filter data using uncertainty-based metrics to isolate challenging samples and then apply clustering or submodular selection to ensure representative coverage of the input space [16]. This approach mitigates the weaknesses inherent in relying solely on one criterion (e.g., redundancy in uncertainty-based selection or neglecting difficult cases in pure diversity-based methods), leading to a balanced and informative training set. The general mathematical formulation of hybrid selection defines it as a weighted sum of multiple metrics (most commonly uncertainty and diversity). For categorization in this paper, we also consider those metrics as Hybrid, which do not directly fall under uncertainty or diversity-based methods.

This taxonomy highlights the versatility of pruning methods, with each category suited to different speech processing challenges. For example, uncertainty-based methods excel at identifying informative samples for active learning, while diversity-based approaches ensure broad coverage of acoustic variations. Hybrid methods combine these strengths to create robust and efficient training sets. Examples of the data selection method for each task and category are shown in Fig. 3.

Furthermore, we systematically categorize the data selection methods from the 153 reviewed papers according to three

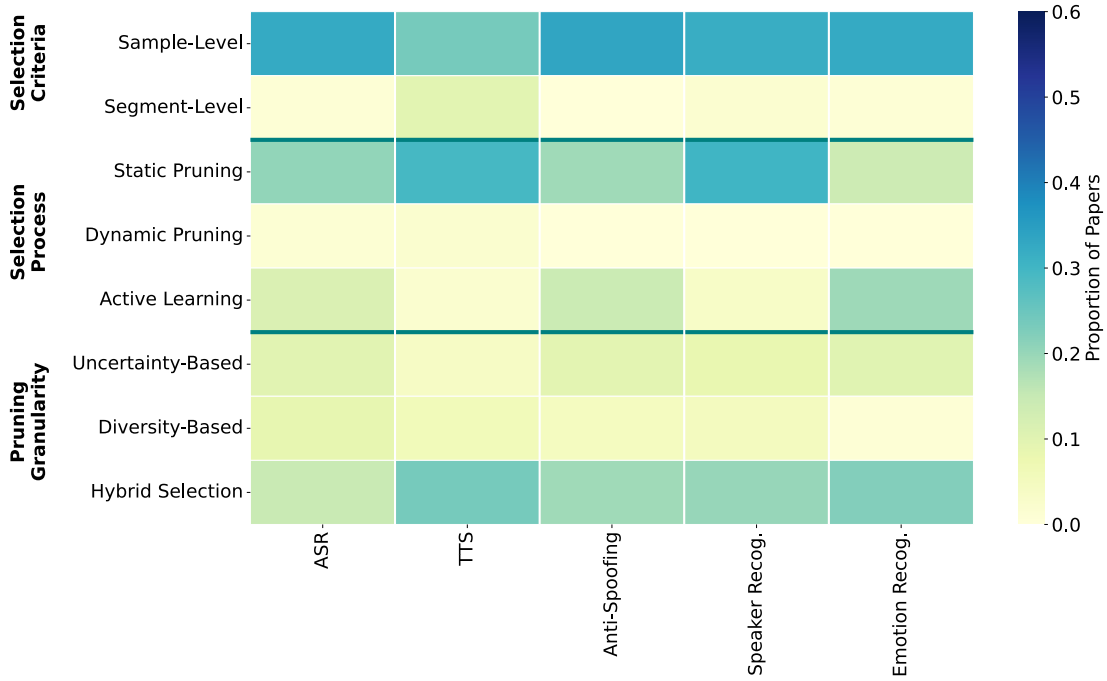

FIGURE 4. Heatmap of 153 papers on data selection techniques in speech processing. This visualization shows the proportion of papers using different data selection approaches (y-axis) across speech processing tasks (x-axis). Techniques are categorized based on pruning granularity, selection process, and selection criteria. Color intensity indicates the fraction of papers using a technique out of the total papers for a particular speech task and category.

key dimensions: pruning granularity, selection process, and selection criteria. The heatmap in Fig. 4 illustrates the distribution of papers across different speech tasks and taxonomy categories. Among these, sample-level selection emerges as the predominant choice across most speech-related tasks, while segment-level selection is notably prevalent in text-to-speech (TTS) applications. Regarding the selection process, active learning methods are predominantly used for emotion recognition tasks, whereas static pruning methods are more frequently adopted for other speech-related tasks. Examining the selection criteria, automatic speech recognition (ASR) research exhibits a balanced usage of uncertainty-based, diversity-based, and hybrid selection approaches. In contrast, for other tasks, hybrid (or other) selection metrics are more commonly favored.

The presented taxonomy and the associated criteria are also applicable in unsupervised and self-supervised learning settings. For example, diversity-based selection can be used to select informative data and reduce the computational requirements for large-scale self-supervised pre-training of speech SSL models.

In the following sections, we explore specific speech tasks in detail and discuss how the selection strategies are adapted and optimized for different objectives.

V. DATA SELECTION FOR SPEECH TASKS

A. AUTOMATIC SPEECH RECOGNITION

1) TASK DEFINITION

Automatic Speech Recognition (ASR) involves converting spoken language into written text. Data selection in ASR [14], [15], [17], [19], [20], [21], [22], [23], [25], [26], [27], [28], [29], [30], [31], [32], [34], [35], [36], [37], [38], [39], [41], [42], [43], [44], [45], [46], [47], [49], [50], [51], [52], [53], [54], [56], [57], [58], [59], [60], [61], [64], [65], [66], [67], [68], [70], [71], [72], [73], [74], [76], [77], [78], [79], [80], [81], [83], [84], [85], [86], is particularly crucial as the training data must cover diverse phonetic contexts, accents, speaking styles, and acoustic conditions to ensure robust recognition performance. The task requires careful consideration of phonemic richness and word coverage to build models that can generalize across different speakers and environments. This becomes especially important in the context of modern self-supervised learning approaches, where the quality and representativeness of the training data directly impact the model's ability to learn meaningful speech representations. Table 1 provides a summary of key data selection methods discussed in this section.

2) CURRENT RESEARCH

a: DIVERSITY-BASED METHODS

Existing work on data selection for ASR has evolved from simple sampling strategies to sophisticated optimization techniques. Early research highlighted the necessity of using phonemically rich text and achieving extensive word coverage. For instance, Wu et al. [26] demonstrated that uniformly sampling a subset across phonemes and words outperforms random selection. Similarly, Kleynhans and Barnard [15] found that selecting utterances that match the target frequency distribution improved limited-data ASR

TABLE 1. Categorization and performance of key data selection methods in ASR.

Study	Selection Process	Pruning Granularity	Selection Criteria	Performance
Just et al. [17]	Static	Sample-level	Hybrid	13-17% relative WER reduction over heuristic baseline under fixed data budgets (40-100h)
Azeemi et al. [16]	Static	Sample-level	Diversity	Up to 17% relative WER improvement over baselines with significant efficiency improvements
Hakkani-Tür et al. [18]	Active Learning	Sample-level	Uncertainty-based	27% reduction in required labeled data to achieve a given accuracy
Wei et al. [14]	Static	Sample-level	Diversity-based	Achieved comparable performance using only 5% of training data
Kothawade et al. [19]	Static	Sample-level	Diversity-based	3-5× higher label efficiency than competing methods
Riccardi et al. [20]	Active Learning	Sample-level	Uncertainty-based	Reduced labeled data needs by >60% vs. random sampling
Yu et al. [21]	Active Learning	Sample-level	Uncertainty-based	Outperforms confidence-based methods; reduced data needs by 50% vs. confidence-based and 60% vs. random sampling for same accuracy
Kundacina et al. [22]	Active Learning	Sample-level	Hybrid	Effective selection of diverse and informative samples
Dossou et al. [23]	Active Learning	Sample-level	Uncertainty-based	27% relative WER improvement with 45% less data vs. baselines
Xiao et al. [11]	Dynamic	Segment-level	Diversity-based	Up to 1.6x training time reduction with negligible performance loss
Gody et al. [24]	Static	Sample-level	Hybrid	Improved performance in limited transcription budget
Wang et al. [25]	Static	Sample-level	Uncertainty-based	14.0-14.7% relative CER reduction

compared to random selection. Just et al. [17] leveraged neural scaling laws to estimate how adding data from different sources affects WER, guiding the selection of an optimal mix. In their experiments on LibriSpeech and TED-LIUM, an informed selection showed 13–17% relative WER reduction over a random baseline, under fixed 40–100 hour data budgets.

b: SUBMODULAR OPTIMIZATION APPROACHES

The submodular optimization framework has emerged as a particularly effective approach for ASR data selection. Wei et al. [14] addressed the problem of selecting a subset from large acoustic datasets using constrained submodular function maximization, showing that DNN-based systems using only about 5% of the training data could achieve performance comparable to systems trained on complete datasets. Kothawade et al. [19] introduced DITTO (Data-Efficient and Fair Targeted Subset Selection for ASR Accent Adaptation), which uses submodular mutual information to choose a subset that best represents a target accent for accent-adapted ASR. DITTO achieved 3–5× higher label efficiency than other methods on accent-specific ASR, meaning it reached a given accuracy with far less data.

c: ACTIVE LEARNING

Active learning has emerged as another powerful approach for ASR [33], [35], [36], [38], [40], [41], [42], [45], [47], [49], [51], [58], [63], [67], [70], [72], [73], [76], particularly in low-resource settings where transcription is expensive. Riccardi and Hakkani-Tur [20] introduced an online algorithm that estimates confidence scores through lattice outputs to select the most informative utterances, reducing labeled data needs by over 60% compared to random sampling. Similarly, Hakkani-Tür et al. [18] demonstrated a 27% reduction in required labeled data through confidence-based selective sampling. Yu et al. [21] proposed a unified Global Entropy Reduction Maximization (GERM) framework that outperforms traditional confidence-based approaches for both active learning and semi-supervised learning. Kundacina et al. [22] developed a two-stage pipeline combining unsupervised (x-vectors clustering) and supervised strategies (Bayesian AL with Monte Carlo dropout) for diverse and informative sample selection. For targeted applications, Dossou [23] introduced an epistemic uncertainty-driven selection approach for African-accented ASR, achieving 27% WER improvement while requiring 45% less data than established baselines. When comparing methodologies for low-resource languages, Syed et al. [27] found that well-designed unsupervised methods can perform nearly as well as supervised approaches.

d: UNSUPERVISED SELECTION METHODS

In parallel, unsupervised data selection has proven valuable for ASR fine-tuning. Gody and Harwath [24] investigated selection techniques for fine-tuning HuBERT models under limited transcription budgets, analyzing the impact of speaker

and topic diversity on performance. Taking a different approach, Lagos and Calapodescu [28] proposed extracting self-supervised representations of both text and audio to create domain-calibrated vector representations, enabling effective data selection through k-nearest neighbor search.

e: SELF-SUPERVISED FINE-TUNING APPROACHES

For fine-tuning self-supervised ASR systems specifically, several innovative approaches have been developed. In the context of accent personalization, Awasthi et al. [29] employed a phoneme-level error model to choose sentences that achieve lower test WER compared to random sentence selection. Similarly, Azeemi et al. [16] proposed COWERAGE, which uses early WER metrics to select diverse training examples. When evaluated on TIMIT, Librispeech, and LJSpeech, COWERAGE achieved up to 17% WER improvement over baselines and demonstrated transferability across SSL models.

f: SPECIALIZED SCENARIO SELECTION

Beyond general ASR, data selection techniques have also been tailored for specific challenging scenarios. For dysarthric speech recognition, Xiong et al. [30] proposed an utterance-based data selection method based on the entropy of posterior probability, showing nearly 2% absolute improvement for speakers with moderate to severe dysarthria. Addressing the unique challenges of children's speech recognition, Wang et al. [25] demonstrated that speaker embedding similarity-based data selection on TTS-augmented data can achieve relative CER reductions of 14.0% and 14.7% for child conversation and reading tasks. Finally, for multilingual and cross-lingual settings, Chen et al. [31] introduced a selection method using Spoken Language Identification models to enhance ASR for low-resource and zero-resource languages.

3) CHALLENGES AND FUTURE OPPORTUNITIES

Several challenges persist in ASR data selection. One primary challenge is handling the inherent variability in speech data, including different accents, speaking rates, and acoustic environments. Current selection methods must balance between maintaining sufficient diversity while pruning redundant or less informative examples. Additionally, the emergence of self-supervised learning models has introduced new challenges in selecting pre-training data, as these models require different characteristics in training data compared to traditional supervised approaches.

Another major challenge for ASR data selection is domain shift. Models trained on one domain (e.g., audiobooks) degrade on others (e.g., conversational speech). Selecting data similar to the target domain can help, but identifying that similarity may require transcriptions or metadata. Untranscribed audio abundance versus labeled scarcity is another issue; thus, semi-supervised selection using pseudo-labels is being explored recently [32]. Additionally, ensuring speaker and dialect balance is crucial to avoid bias; otherwise, an ASR may perform poorly on under-represented accents. Recent research emphasizes fairness in ASR, selecting data to improve performance on marginalized dialects and reduce demographic WER gaps.

Looking ahead, there are numerous opportunities for advancing ASR data selection methods. The development of more comprehensive metrics for assessing the informativeness of speech segments, particularly in the context of self-supervised learning, represents a promising direction. There is also potential for developing adaptive selection strategies that can dynamically adjust to different domains and languages. The presented taxonomy (granularity, process, criteria) readily applies to multilingual and code-switching scenarios; for instance, segment-level selection can target code-switched portions, while diversity criteria can ensure coverage across languages. The integration of linguistic knowledge with data-driven selection criteria could lead to more efficient and effective training sets. The field would also benefit from more research into how data selection strategies can be optimized for specific deployment scenarios, such as low-resource languages or specialized domains.

B. TEXT-TO-SPEECH SYSTEMS

1) TASK DEFINITION

Text-to-speech (TTS) synthesis involves converting written text into natural-sounding speech, requiring high-quality training data that covers diverse phonetic contexts and prosodic variations. Data selection in TTS [25], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100], [101], [102] is particularly crucial as the training data must encompass a wide range of phonetic contexts to ensure the system can synthesize any input text with high fidelity. The task becomes even more challenging in scenarios involving low-resource languages or when dealing with found data that lacks corresponding text transcriptions. Key methods discussed in this section are summarized in Table 2.

2) CURRENT RESEARCH

a: CORPUS OPTIMIZATION METHODS

Black and Lenzo [88] developed methods for optimizing unit-selection TTS datasets by maximizing phonetic coverage and acoustic diversity. Their greedy approach selects acoustic distinctions to capture variations in data, ensuring high-quality concatenative synthesis. Recent research has increasingly focused on developing TTS systems for low-resource languages by leveraging "found" data such as broadcast news, ASR corpora, and audiobooks. In this context, Tesfaye Biru et al. [89] demonstrated that careful selection of training subsets based on acoustic-prosodic criteria, such as high mean energy, low variability in energy and f0, and the incorporation of gemination information, can significantly enhance both the intelligibility and naturalness of synthesized speech. These findings show the potential of strategic data selection and adaptation in overcoming the challenges associated with scarce annotated speech data for

TABLE 2. Categorization and performance of key data selection methods in TTS.

Study	Selection Process	Pruning Granularity	Selection Criteria	Performance
Baali et al. [87]	Static	Sample-level	Hybrid	More natural speech (MOS Intelligibility: 4.4, Naturalness: 4.2) from smaller, filtered datasets
Black et al. [88]	Static	Segment-Level	Hybrid	Improved quality in unit-selection TTS
Tesfaye et al. [89]	Static	Sample-level	Hybrid	Enhanced intelligibility and naturalness for low-resource TTS
Kuo et al. [90]	Static	Sample-level	Hybrid	More natural speech output from smaller, curated dataset
Chalamandaris et al. [91]	Static	Segment-Level	Hybrid	Optimal corpus design for high-quality Bulgarian TTS
Zhao et al. [92]	Active Learning	Segment-Level	Uncertainty-based	Improved boundary prediction with reduced annotation
Seki et al. [93]	Dynamic	Sample-level	Uncertainty-based	Superior performance vs. conventional acoustic-quality methods
Jung et al. [94]	Static	Sample-level	Hybrid	Demonstrated feasibility of TTS with less controlled data
Braunschweiler et al. [95]	Static	Sample-level	Diversity-based	Significant improvements over unfiltered corpus
Cooper et al. [96]	Static	Sample-level	Hybrid	More natural-sounding voices through strategic pruning
Seki et al. [97]	Static	Sample-level	Diversity-based	Outperformed baseline phoneme-balanced selection
Mandeel et al. [98]	Static	Sample-level	Hybrid	Successfully modeled creaky voice characteristic
Gonzalez et al. [99]	Static	Sample-level	Hybrid	Significantly improved quality for low-quality corpora; 3h selected data comparable to much larger datasets

low-resource language TTS systems. Kuo et al. [90] explore data selection for enhancing the naturalness in TTS voices trained on found corpora that are small in size. Their work focuses on systematically rejecting segments that degrade synthesis quality by introducing metrics such as the Phone Matching Error Rate (PMER), voiced/unvoiced mismatch, Signal-to-Noise Ratio (SNR), and articulation measures. Experiments on Mandarin audiobook datasets demonstrated that training on a smaller, carefully curated subset can lead to more natural-sounding speech than using the complete dataset.

b: CORPUS DESIGN

Chalamandaris et al. [91] addressed corpus design for unit selection TTS in the context of Bulgarian. By systematically selecting sentences that cover a target set of diphones and refining the corpus through synthesis-driven evaluation, they illustrated that an optimally designed corpus is pivotal for high-quality TTS. Their strategy highlights the advantage of a multi-stage approach for increasing diphone and unit coverage, advocating for a tailored approach to corpus design. Yu et al. [103] presented a data pruning method for concatenative Chinese TTS systems based on the segmentation of syllables using three portions. Different factorial acoustic features are used, and the pruning process removes bad syllables.

c: ACTIVE LEARNING AND DYNAMIC SELECTION

Zhao and Ma [92] pioneered the application of TTS active learning for predicting prosodic word boundaries in Chinese TTS using Maximum Entropy Markov Models. By selectively choosing the most informative samples for annotation, they demonstrated improved boundary prediction with reduced annotation effort. Subsequently, Seki et al. [93] proposed an evaluation-in-the-loop data selection method that selects training data based on the predicted quality of synthetic speech rather than the acoustic quality of the original recordings. Their approach demonstrated superior performance compared to conventional acoustic-quality-based methods when tested on YouTube data, highlighting the importance of considering the end goal of TTS in the selection process.

d: UNCONVENTIONAL DATA SOURCES

The challenge of selecting data from unconventional sources has received significant attention. Jung et al. [94] introduced the TTS In the Wild (TITW) dataset derived from VoxCeleb1, establishing benchmarks for training TTS systems with data collected from less-controlled environments. Their work demonstrates that while carefully selected and enhanced subsets (TITW-Easy) can produce acceptable results, using challenging data (TITW-Hard) remains difficult for current models. Similarly, Baali et al. [87] proposed an unsupervised method for data selection from broadcast news for Arabic TTS, showing that smaller, carefully selected datasets can produce more natural speech than larger, unfiltered datasets.

e: AUDIOBOOK DATA SELECTION

Another significant contribution comes from Braunschweiler and Buchholz [95], who investigated automatic sentence selection from speech corpora to improve HMM-based TTS

synthesis quality. Focusing on audiobook data, their research tackles the challenges posed by diverse speaking styles and non-neutral speech. By combining acoustic features (e.g., f0, RMS amplitude, and voicing) and text based cues (such as punctuation and normalization error indicators), they developed an automatic method that discards problematic sentences. Although manual selection of 'neutral' speech still showed the best performance, their automatic approach provided significant improvements over using the full unfiltered corpus, demonstrating a viable route for large-scale TTS corpus preparation. Cooper et al. [96] further contribute to this body of work by exploring data selection and adaptation for naturalness in HMM-based speech synthesis using broadcast news data. Their experiments compared various selection strategies, such as removing hyper-articulated utterances and favoring low mean f0 and hypo-articulated speech, to optimize the naturalness of the synthesized output. Evaluation via crowdsourced listening tests confirmed that carefully pruning outlier utterances enhances overall voice quality. This work not only reinforces the benefit of strategic data selection but also highlights the potential for adapting these techniques to multiple speakers and diverse recording conditions, an important consideration for low-resource language scenarios.

f: DIVERSITY-BASED METHODS

Recent work has expanded the focus to diversity-based methods. Seki et al. [97] proposed a diversity-based core-set selection approach that uses linguistic and acoustic features to ensure wide coverage with minimal data. Their method outperformed baseline phoneme-balanced selection across different languages and the size of the corpus.

g: VOICE MODELING

For scenarios with specialized voice characteristics, Mandeel et al. [98] focused on modeling irregular voice in end-to-end speech synthesis through strategic speaker adaptation. By choosing the data based on the presence of irregular (or *creaky*) phonation, they successfully modeled this voice characteristic, contributing to more personalized speech synthesis. González-Docasal and Álvarez [99] enhanced voice cloning quality through data selection, demonstrating significant improvements for challenging low-quality corpora through strategic data ablation.

h: HANDLING UNTRANSCRIBED AUDIO

In scenarios where audio data lacks corresponding text, Godambe et al. [101] addressed this issue by proposing a method to build unit selection voices from such data. Their approach includes using a large vocabulary public domain ASR system to generate transcriptions from the audio, then applying confidence measure-based data pruning to remove problematic segments. The effectiveness of their method was evaluated through perceptual listening tests, comparing synthesized speech from pruned data to that from manually transcribed data.

C. CHALLENGES AND FUTURE OPPORTUNITIES

Several significant challenges persist in TTS data selection. Handling noisy and variable data in found corpora requires more sophisticated selection techniques that can effectively distinguish between usable and problematic segments. The reliance on ASR for transcription in data without text can introduce errors, necessitating improved confidence measures and more robust error detection methods. Additionally, the scalability of these methods to very large datasets and their generalization across languages need further exploration.

Looking ahead, future research opportunities lie in developing machine learning models for automatic data selection, potentially using deep learning to identify optimal subsets based on acoustic and prosodic features. DL models could predict synthesis quality or use embeddings for better diversity metrics, though applying these complex models at scale presents computational challenges. There is also a need for robust quality assessment metrics that can automatically evaluate the suitability of speech segments for TTS training. The emergence of multilingual TTS systems presents new challenges in selecting representative data across different languages while maintaining consistent quality. Data selection is crucial here, especially for low-resource languages: diversity criteria can maximize phonetic/prosodic coverage from limited corpora, active learning can guide targeted data acquisition and cross-lingual selection can identify relevant data from high-resource languages for transfer learning. Finally, the integration of emotional and stylistic variations in TTS training data selection remains an important area for investigation as modern systems increasingly aim to produce more expressive and context-appropriate speech synthesis.

D. AUDIO ANTI-SPOOFING

1) TASK DEFINITION

Audio anti-spoofing (or speaker verification countermeasures) involves distinguishing genuine speech from spoofed speech attacks (synthetic speech, voice conversion, or replay recordings). Data selection in audio-antispoofing [104], [105], [106], [107], [108], [109], [110], [111] refers to choosing a variety of bona fide and spoof samples that prepare the system to detect even unknown attacks. A key challenge is that spoofing can be done through various methods (different text-to-speech engines, voice conversion techniques, replay attacks), so the training data needs to cover a broad range of attack types. Table 3 highlights several important data selection approaches for audio anti-spoofing.

2) CURRENT RESEARCH

a: SUPERVISED PRUNING METRICS

Azeemi et al. [104] introduced supervised metrics for pruning spoof datasets. Their method, which includes metrics like the Forgetting Norm, combines error magnitude with forgetting events to identify persistently challenging examples. This approach achieved a 23% relative improvement in equal

TABLE 3. Categorization and performance of key data selection methods in audio anti-spoofing.

Study	Selection Process	Pruning Granularity	Selection Criteria	Performance
Azeemi et al. [104]	Static	Sample-level	Uncertainty-based	Up to 23% relative EER improvement (using forgetting norm) over baseline heuristics on ASVspoof 2019
Azeemi et al. [105]	Static	Sample-level	Diversity-based	Exceeded performance of 4 other metrics; 91% faster computation (vs. metrics requiring initial training)
Dhamyal et al. [106]	Static	Sample-level	Hybrid	Effective lightweight indicators; EER of 23.5% using VOT+coarticulation fusion
Sanchez et al. [107]	Static	Sample-level	Hybrid	Notable EER improvements
Jimenez et al. [108]	Active Learning	Sample-level	Hybrid	Improved detection error rates with fewer training examples
Wang et al. [109]	Active Learning	Sample-level	Hybrid	>40% relative reduction in detection error rates

error rate (EER) over baseline methods on the ASVspoof 2019 dataset. Subsequent work by Azeemi et al. [105] introduced label-free pruning using audio embeddings. This self-supervised approach leverages wav2vec2 embeddings with k-means clustering to identify prototypical examples, resulting in a 91% faster computation compared to supervised methods by eliminating the initial training phase. The analysis also revealed differential preservation of attacks in pruned subsets.

b: FEATURE-DRIVEN STRATEGIES

Complementary research has explored alternative feature-driven and data-centric strategies for audio anti-spoofing. Dhamyal et al. [106] propose leveraging microfeatures, specifically Voicing Onset Time (VOT) and articulation, to distinguish between genuine and synthesized speech. Their study shows that these fine-grained acoustic measurements, which capture subtle differences in speech production, can serve as effective, lightweight indicators for spoof detection, particularly in resource-constrained environments. Meanwhile, Sanchez Valera et al. [107] focus on data augmentation techniques for Physical Access (PA) replay attacks. They demonstrate that methods such as time masking, additive noise, Room Impulse Response filtering, and mixup can significantly enhance the robustness of anti-spoofing models. Their LCNN-based classifier, trained on augmented versions of the ASVspoof 2019 corpus, exhibits notable improvements in EER, showing the value of synthetic diversity to bridge the gap between simulated and real-world attack scenarios.

c: ACTIVE DATA SELECTION

Active data selection has also emerged as a promising avenue for refining training sets for audio anti-spoofing. Jiménez Garizao [108] implement several active learning (AL) algorithms that iteratively select the most informative samples from a large candidate pool for training anti-spoofing countermeasures (CMs). Their work achieved significant improvements in the detection error rates while using fewer training examples than conventional methods. The methodology was tested on ASVSpoof 2019 (English) and the Habla (Spanish) datasets. In another work on active learning, Wang and Yamagishi [109] investigate active-learning-based data selection for speech spoofing countermeasures. Their study refines the AL framework by not only selecting challenging data but also actively removing uninformative samples from the training pool. This dual strategy further improves the generalization of the CM across diverse attack conditions, and they show that this strategy achieves a relative reduction of more than 40% error detection rates.

3) CHALLENGES AND FUTURE OPPORTUNITIES

Generalization to new attacks is a core challenge in audio anti-spoofing. As the ASVSpoof 2021 results showed [112], systems that performed well on known attack types often failed when faced with different conditions. Data selection must encompass variability: microphone types, noise backgrounds, speakers, and languages (mostly English so far, a gap for low-resource languages [113], [114]). Another challenge is data scarcity for certain spoof types – e.g., only a few deepfake audio examples might be available publicly, making it hard to train detection. The community often must generate spoofs (using TTS/VC models) to enlarge training sets, essentially simulating the adversary. Ensuring these simulated attacks are representative of actual threat examples is non-trivial. Label quality is generally straightforward (genuine vs spoof), but in replay data, sometimes labeling segments (if part of an utterance is replay vs live) can be complex. Evolving attacks mean the *optimal* training set is a moving target; ongoing data collection (like adding samples of new voice conversion methods) is needed.

Quality control in data also remains a significant challenge: anti-spoofing models can inadvertently learn irrelevant cues if the dataset isn't cleaned. For example, if all spoof recordings have a certain background hum or all genuine recordings are high quality, the model might latch onto noise level instead of true speech artifacts. Future work might leverage deep learning to automatically identify subtle spoofing artifacts or predict hard examples for selection, though the computational cost of applying such methods to large datasets requires careful consideration regarding scalability.

TABLE 4. Categorization and performance of key data selection methods in speaker recognition.

Study	Selection Process	Pruning Granularity	Selection Criteria	Performance
Lei et al. [115]	Static	Sample-level	Hybrid	2.13% EER improvement at the fifth of the computational costs
Abdullah et al. [116]	Static	Sample-level	Hybrid	Achieved high accuracy (e.g., 97.3% speaker rec.) while reducing features via GA
Huang et al. [117]	Static	Sample-level	Uncertainty-based	Improved performance on NIST speaker recognition corpus
Pullella et al. [118]	Static	Sample-level	Hybrid	Significant improvement in noisy conditions
Sun et al. [119]	Static	Sample-level	Hybrid	Effective selection of high-quality frames while discarding corrupted ones
Nath et al. [120]	Static	Sample-level	Hybrid	Formant+LPC feature leads to high speaker recognition rate

E. SPEAKER RECOGNITION

1) TASK DEFINITION

Speaker recognition involves identifying or verifying a person's identity from their voice characteristics. The task requires processing speech data to extract speaker-specific features while being robust to variations in recording conditions, speaking style, and content. Data selection in this context [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126], [127], [128], [129], [130], [131], [132], [133], [134] focuses on choosing the most representative and discriminative speech segments that capture unique speaker characteristics while minimizing redundancy and noise. A summary of the discussed speaker recognition data selection methods can be found in Table 4.

2) CURRENT RESEARCH

a: MAXIMUM ENTROPY APPROACH

Huang and Ma [117] pioneered a data selection approach based on a maximum entropy criterion to prune training data for speaker recognition. Working within the GMM-UBM framework (Gaussian mixture model – Universal Background Model), they remove redundant feature frames when training the UBM and select only the most informative frames for each target speaker model. This approach ensures models focus on distinctive speaker characteristics rather than being overwhelmed by the sheer quantity of frames, resulting in better performance compared to the baseline on the NIST speaker recognition corpus.

b: NOISE-ROBUST SELECTION

Researchers have attempted to enhance the robustness of speaker recognition systems under noisy conditions. Pullella et al. [118] proposed a method that combines missing data processing with dynamic feature selection. Their technique refines time-frequency reliability masks by eliminating non-discriminative frequency sub-bands based on speaker-specific information. By integrating bottom-up noise estimation with top-down discriminative feature analysis (using multi-condition training and SNR estimates), their system is able to significantly improve speaker identification performance over traditional processing.

c: NEURAL NETWORK FEATURE SELECTION

In a parallel line of research, Sun et al. [119] developed an efficient feature selection method for speaker recognition using a neural network framework. Their study investigates various spectral features to identify the optimal strategy for robust speaker modeling that keeps high-quality frames and discards the noisier or corrupted frames. Notably, their experiments demonstrate that the spectral subtraction-based method can select efficient feature frames effectively, highlighting the importance of such selection criteria in noisy environments.

d: ADVANCED FEATURE EXTRACTION

Furthermore, Nath and Kalita [120] explored feature extraction and selection using feed-forward neural networks. By comparing different feature combinations among formants, LPCC, and MFCC and analyzing their impact on recognition performance, they reinforced the conclusion that carefully chosen features improve accuracy. Together, these contributions underline the current research trend toward integrating intelligent data and feature selection strategies to develop speaker recognition systems that are both robust and efficient.

3) CHALLENGES AND FUTURE OPPORTUNITIES

A key challenge in speaker recognition data selection is ensuring balanced representation across different recording conditions and speaking contexts. Current systems must carefully manage the trade-off between having sufficient data per speaker and avoiding bias towards particular recording sessions or acoustic environments. Ensuring a balanced subset of speech from each target speaker (not too much from any one session) remains a practical consideration to avoid biasing towards one channel or context.

Future research opportunities include developing more sophisticated methods for identifying truly distinctive speaker characteristics in the presence of channel effects and environmental noise. Additionally, there is potential

for exploring adaptive data selection strategies that can dynamically adjust to different speaker verification scenarios and deployment conditions.

F. EMOTION RECOGNITION

1) TASK DEFINITION

Speech emotion recognition aims to automatically identify emotional states from speech signals. The task is particularly challenging due to the subjective nature of emotions, the variability in their expression across speakers and cultures, and the often subtle acoustic cues that distinguish different emotional states. Data selection in emotion recognition [135], [136], [137], [138], [139], [140], [141], [142], [143], [144], [145], [146], [147], [148], [149], [150], [151], [152], [153], [154], [155], [156], [157], [158], [159], [160], [161], [162], [163], [164], [165], [166], [167], [168] is crucial because not all recorded utterances carry strong emotional content, and some may have ambiguous labels, making the quality and reliability of training data particularly important. Table 5 summarizes key data selection techniques for speech emotion recognition.

2) CURRENT RESEARCH

a: SUB-UTTERANCE SELECTION

Le et al. [135] pioneered sub-utterance selection strategies by identifying and isolating those segments within an utterance that are most emotionally salient. By focusing on high-emotion segments (e.g., parts where excitement or sadness peaks), their approach prunes out low-relevance sections and allows emotion classifiers to be trained more efficiently. The results also highlight that the sub-utterance strategy leads to more stable learning and slightly faster convergence, highlighting the benefits of segment / sub-utterance level selection.

b: ACTIVE LEARNING STRATEGIES

More recent work has focused on reducing annotation effort and improving generalization through active learning (AL) strategies [143], [144], [146], [148], [152], [154], [157], [158], [161], [162], [165], [166], [167]. Several studies have adapted AL for both classification and regression in dimensional SER. For example, Han et al. [136] applied the SVM-based AL method to select the most informative unlabeled speech instances based on uncertainty measures, achieving competitive performance with a significantly reduced labeling budget. Similarly, Zhao and Ma [137] developed an AL framework using Conditional Random Fields to actively select samples for annotation, demonstrating that AL can match or exceed the performance of random sampling strategies.

c: DOMAIN ADAPTATION TECHNIQUES

Other researchers have combined AL with domain adaptation techniques. Abdelwahab and Busso [138] proposed an incremental adaptation algorithm that combines AL and supervised domain adaptation to selectively annotate target domain samples for acoustic emotion recognition, thereby incrementally refining the classifier with only the most confidently correct samples. In a subsequent work, Li et al. [139] introduced the AFTER framework, which integrates task adaptation pre-training with AL to fine-tune large-scale pre-trained ASR models for SER, resulting in notable gains in both accuracy and efficiency simultaneously. Finally, Ren et al. [140] presented an integrated active learning framework for constructing speech emotion corpora. Their method prioritizes the selection of samples from underrepresented emotion classes, effectively addressing dataset imbalance and reducing labeling costs. Collectively, these active learning approaches not only reduce the annotation burden but also lead to more robust and efficient SER systems.

3) CHALLENGES AND FUTURE OPPORTUNITIES

A major challenge in emotion recognition is dealing with highly imbalanced and limited datasets. Current selection techniques are crucial for maximizing the information content of training sets and handling class imbalance by selectively oversampling under-represented emotion categories or undersampling neutral/dominant classes. However, determining the optimal balance between different emotional categories while maintaining natural distribution remains an open problem. Data selection for speech emotion recognition must balance between quality, quantity, and realism. Previous works have also shown a trade-off: acted datasets have strong signals but are limited and somewhat exaggerated, while real datasets are scarce and often noisy. A combination might be the best path, along with techniques to expand data (augmentation, transfer learning, active learning) to cover what cannot be directly selected due to lack of availability.

G. OTHER SPEECH TASKS

1) KEYWORD SPOTTING

Keyword spotting (KWS) involves detecting specific words or phrases in audio streams [170], a task central to voice-activated systems and speech indexing, yet often constrained by limited training data in low-resource scenarios. Fraga-Silva et al. [12] investigate active learning (AL) for selecting training data from an untranscribed pool for manual transcription, targeting improved speech-to-text (STT) and KWS systems within the IARPA-Babel program. Their approach compares AL-selected data against a predefined Very Limited Language Pack (VLLP) baseline. For Lithuanian, their AL methods reduced WER by 1-1.7% absolute and improved Actual Term Weighted Value (ATWV) by approximately 3% absolute over the VLLP baseline.

2) SPEAKER DIARIZATION

Speaker diarization, the task of identifying *who spoke when* in a speech signal, is a critical component of speech processing, particularly for applications like automated analysis and transcription [171]. Yu and Hansen [169] introduce an active

TABLE 5. Categorization and performance of key data selection methods in speech emotion recognition.

Study	Selection Process	Pruning Granularity	Selection Criteria	Performance
Le et al. [135]	Static	Segment-Level	Diversity	Achieves median unweighted average recall of 70.68% using less than 50% of the training data.
Han et al. [136]	Active Learning	Sample-level	Uncertainty-based	Up to 12% reduction in training data for achieving certain performance
Zhao et al. [137]	Active Learning	Sample-level	Uncertainty-based	Exceeded random sampling performance
Abdelwahab et al. [138]	Active Learning	Sample-level	Hybrid	Significant improvements over conventional adaptation schemes via carefully selected samples
Li et al. [139]	Active Learning	Sample-level	Hybrid	Achieved +8.45% accuracy and 79% less time using only 20% selected data
Ren et al. [140]	Active Learning	Sample-level	Hybrid	Reduced labeling costs; achieved 90% accuracy with <50% labeled data
Lin et al. [141]	Static	Sample-level	Hybrid	Enhanced cross-lingual generalization
Li et al. [142]	Static	Sample-level	Hybrid	Improved estimation of emotion dimensions (lower MAE, higher correlation) via feature selection

TABLE 6. Categorization and Performance of key data selection methods in other speech tasks.

Study	Selection Process	Pruning Granularity	Selection Criteria	Performance
Fraga-Silva et al. [12]	Active Learning	Sample-level	Uncertainty-based	Reduced WER by 1-1.7% absolute over VLLP and 3% absolute over ATWV
Yu et al. [169]	Active Learning	Sample-level	Diversity-based	Significantly reduced diarization error rates

learning-based constrained clustering approach for speaker diarization designed to optimize performance with minimal human input. Their method proceeds in two stages: an exploratory phase that identifies initial speaker samples to form reliable clusters, followed by constrained clustering, where these initial clusters are preserved while further refining the segmentation. To maximize efficiency, they also propose selecting speech segments with the highest expected speaker error for human review, enhancing cluster assignments iteratively. Evaluated on the Apollo Mission Control Center dataset and augmented multiparty meeting corpora, their approach significantly reduces diarization error rates with modest human effort.

Table 6 summarizes the discussed methods for keyword spotting and speaker diarization.

VI. CONCLUSION

This survey has demonstrated that data selection techniques have enabled efficient speech processing across various tasks – emotion recognition, speaker recognition, speech enhancement, automatic speech recognition, and beyond. The key insight that emerges across these domains is that data quality and relevance frequently outweigh sheer quantity. Intelligent data selection strategies accomplish several critical objectives: they address class imbalance problems (particularly evident in emotion recognition tasks), eliminate redundant or noisy samples (as shown in speaker recognition systems), and better match training data to target conditions (demonstrated in speech recognition applications). By identifying and focusing on the most valuable portions of datasets while pruning away less useful samples, researchers have consistently achieved models that are more efficient, more robust, and in many cases, more accurate than those trained on unfiltered datasets.

Our taxonomic organization of data pruning methods has revealed the diversity of approaches. Each method offers distinct advantages depending on the specific speech task, dataset characteristics, and computational constraints. This systematic categorization provides researchers with a framework to select appropriate pruning strategies for their particular applications.

Despite the promising results shown by current data pruning methods, significant challenges remain. Ensuring the scalability of these techniques to ever-larger datasets, maintaining robustness across diverse speech conditions, and guaranteeing fairness by quantifying and preventing the amplification of biases in pruned datasets are critical areas requiring further investigation. The risk of over-pruning, especially in low-resource settings, requires careful consideration to avoid losing critical information and degrading generalization. Furthermore, evaluating the computational complexity of different selection methods presents a challenge, as reporting varies widely in the literature. Generally, static methods incur upfront costs but speed up training, while dynamic and active learning methods add runtime overhead dependent on the calculation frequency and complexity of the

selection criteria. Additionally, while most existing research has focused on supervised learning scenarios, extending data pruning to self-supervised and semi-supervised learning paradigms represents an important frontier for future work. As the computational resources required for state-of-the-art speech models continue to grow, intelligent data selection offers a path toward more sustainable AI development. By reducing training data requirements without sacrificing performance, these techniques can help democratize access to speech technology.

A. FUTURE RESEARCH DIRECTIONS

Based on our analysis, we identify several promising directions for future research in data selection for speech processing:

- **Cross-task transferability:** Investigate adapting effective selection techniques from one speech task (e.g., ASR) to others (e.g., TTS), promoting generalizable methodologies.
- **Automated selection criteria:** Utilize deep learning methods to automatically discover selection criteria from data, moving beyond heuristics.
- **Fairness-aware selection:** Develop dataset selection methods that explicitly address fairness and bias mitigation.
- **Reinforcement learning-based data selection:** Explore reinforcement learning to dynamically learn optimal data selection policies based on model feedback and long-term objectives.
- **Resource-efficient selection:** Design lightweight and computationally efficient selection strategies that minimize the overhead associated with both data selection and subsequent model training.

Addressing these directions can advance efficient, accessible, and sustainable speech technology development, reinforcing the importance of data selection for balancing performance and computational needs.

REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2022, pp. 28492–28518.

[3] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, "VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers," 2024, *arXiv:2406.05370*.

[4] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020, *arXiv:2001.08361*.

[5] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 3645–3650. [Online]. Available: https://aclanthology.org/P19-1355

[6] B. R. Bartoldson, B. Kailkhura, and D. Blalock, "Compute-efficient deep learning: Algorithmic trends and opportunities," *J. Mach. Learn. Res.*, vol. 24, pp. 1–77, Jan. 2022.

[7] R. Taheri, M. Ahmadzadeh, and M. R. Kharazmi, "A new approach for feature selection in intrusion detection system," *Fen Bilimleri Dergisi*, vol. 36, no. 6, pp. 1–10, 2015.

[8] B. Takkouche and G. Norman, "PRISMA statement," *Epidemiology*, vol. 22, no. 1, p. 128, 2011.

[9] B. Bergsma, M. Brzezinska, O. V. Yazyev, and M. Cernak, "Cluster-based pruning techniques for audio data," 2023, *arXiv:2309.11922*.

[10] Y. Lin, Y. Fu, J. Zhang, Y. Liu, J. Zhang, J. Sun, H. H. Li, and Y. Chen, "SpeechPrune: Context-aware token pruning for speech information retrieval," 2024, *arXiv:2412.12009*.

[11] Q. Xiao, P. Ma, A. Fernandez-Lopez, B. Wu, L. Yin, S. Petridis, M. Pechenizkiy, M. Pantic, D. C. Mocanu, and S. Liu, "Dynamic data pruning for automatic speech recognition," 2024, *arXiv:2406.18373*.

[12] T. Fraga-Silva, J.-L. Gauvain, L. Lamel, A. Laurent, V.-B. Le, and A. Messaoudi, "Active learning based data selection for limited resource STT and KWS," in *Proc. Interspeech*, Sep. 2015, pp. 3159–3163.

[13] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, Y. Shiga, and A. Ito, "Entropy-based sentence selection for speech synthesis using phonetic and prosodic contexts," in *Proc. Interspeech*, Sep. 2015, pp. 3491–3495.

[14] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes, "Submodular subset selection for large-scale speech training data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3311–3315.

[15] N. T. Kleynhans and E. Barnard, "Efficient data selection for ASR," *Lang. Resour. Eval.*, vol. 49, no. 2, pp. 327–353, Jun. 2015.

[16] A. H. Azeemi, I. A. Qazi, and A. A. Raza, "Representative subset selection for efficient fine-tuning in self-supervised speech recognition," 2022, *arXiv:2203.09829*.

[17] H. A. Just, I.-F. Chen, F. Kang, Y. Zhang, A. K. Sahu, and R. Jia. (2023). *ASR Data Selection From Multiple Sources: A Practical Approach on Performance Scaling*. [Online]. Available: https://www.amazon.science/publications/asr-data-selection-from-multiple-sources-a-practical-approach-on-performance-scaling

[18] D. Hakkani-Tür, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, May 2002, pp. IV-3904–IV-3907.

[19] S. Kothawade, A. Mekala, D. C. S. H. Havya, M. Kothyari, R. Iyer, G. Ramakrishnan, and P. Jyothi, "DITTO: Data-efficient and fair targeted subset selection for ASR accent adaptation," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, ON, Canada, 2023, pp. 5810–5822.

[20] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 504–511, Jul. 2005.

[21] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Comput. Speech Lang.*, vol. 24, no. 3, pp. 433–444, Jul. 2010.

[22] O. Kundacina, V. Vincan, and D. Miskovic, "Combining X-vectors and Bayesian batch active learning: Two-stage active learning pipeline for speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 33, pp. 1862–1876, 2025.

[23] B. F. P. Dossou, "Advancing African-accented speech recognition: Epistemic uncertainty-driven data selection for generalizable ASR models," 2023, *arXiv:2306.02105*.

[24] R. Gody and D. Harwath, "Unsupervised fine-tuning data selection for ASR using self-supervised speech models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[25] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian, "Towards data selection on TTS data for children's speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6888–6892.

[26] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, May 2007, pp. 562–565.

[27] A. R. Syed, A. Rosenberg, and E. Kislal, "Supervised and unsupervised active learning for automatic speech recognition of low-resource languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5320–5324.

[28] N. Lagos and I. Calapodescu, "Unsupervised multi-domain data selection for asr fine-tuning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 10711–10715.

[29] A. Awasthi, A. Kansal, S. Sarawagi, and P. Jyothi, "Error-driven fixed-budget ASR personalization for accented speakers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7033–7037.

[30] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7424–7428.

[31] J. Chen, C. Chu, S. Li, and T. Kawahara, "Data selection using spoken language identification for low-resource and zero-resource speech recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2024, pp. 1–6.

[32] P. Rangappa, J. Zuluaga-Gomez, S. Madikeri, A. Carofilis, J. Prakash, S. Burdisso, S. Kumar, E. Villatoro-Tello, I. Nigmatulina, P. Motlicek, K. Pandia, and A. Ganapathiraju, "Speech data selection for efficient ASR fine-tuning using domain classifier and pseudo-label filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.

[33] D. Hakkani-Tur, G. Tur, M. Rahim, and G. Riccardi, "Unsupervised and active learning in automatic speech recognition for call classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Jun. 2004, doi: 10.1109/ICASSP.2004.1326014.

[34] E. Alsharhan and A. Ramsay, "Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition," *Lang. Resour. Eval.*, vol. 54, no. 4, pp. 975–998, Oct. 2020, doi: 10.1007/s10579-020-09505-5.

[35] D. Hakkani-Tür, G. Riccardi, and G. Tur, "An active approach to spoken language processing," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 3, pp. 1–31, Oct. 2006, doi: 10.1145/1177055.1177056.

[36] Y. Long, Y. Li, S. Wei, Q. Zhang, and C. Yang, "Large-scale semi-supervised training in deep learning acoustic model for ASR," *IEEE Access*, vol. 7, pp. 133615–133627, 2019, doi: 10.1109/ACCESS.2019.2940961.

[37] A. Rousseau, "XenC: An open-source tool for data selection in natural language processing," *Prague Bull. Math. Linguistics*, vol. 100, no. 1, pp. 73–82, Oct. 2013, doi: 10.2478/pralin-2013-0013.

[38] Y. Yuan, S.-W. Chung, and H.-G. Kang, "Gradient-based active learning query strategy for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2832–2836, doi: 10.1109/ICASSP.2019.8683089.

[39] A. Nagroski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Jun. 2003, pp. 67–72, doi: 10.1109/ASRU.2003.1318405.

[40] B. Varadarajan, D. Yu, L. Deng, and A. Acero, "Maximizing global entropy reduction for active learning in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 4721–4724, doi: 10.1109/ICASSP.2009.4960685.

[41] J. Luo, J. Wang, N. Cheng, and J. Xiao, "Loss prediction: End-to-end active learning approach for speech recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–7, doi: 10.1109/IJCNN52387.2021.9533839.

[42] Y. Zhao, Y. Cao, X. Pan, and X. Xu, "Tibetan language continuous speech recognition based on active WS-DBN," in *Proc. IEEE Int. Conf. Autom. Logistics*, Aug. 2009, pp. 1558–1562, doi: 10.1109/ICAL.2009.5262707.

[43] S. Li, Y. Akita, and T. Kawahara, "Data selection from multiple ASR systems' hypotheses for unsupervised acoustic model training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5875–5879, doi: 10.1109/ICASSP.2016.7472804.

[44] J. Tian, J. Nurminen, and I. Kiss, "Optimal subset selection from text databases," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Jun. 2005, pp. 305–308, doi: 10.1109/ICASSP.2005.1415111.

[45] H. Li, T. Zhang, R. Qiu, and L. Ma, "Grammar-based semi-supervised incremental learning in automatic speech recognition and labeling," *Energy Proc.*, vol. 17, pp. 1843–1849, Jan. 2012, doi: 10.1016/j.egypro.2012.02.321.

[46] D. Seppi and D. V. Compernolle, "Data pruning for template-based automatic speech recognition," in *Proc. Interspeech*, Sep. 2010, pp. 901–904.

[47] Y. Long, H. Ye, Y. Li, and J. Liang, "Active learning for LF-MMI trained neural networks in ASR," in *Proc. Interspeech*, Sep. 2018, pp. 2898–2902.

[48] S. Wotherspoon, W. Hartmann, M. Snover, and O. Kimball, "Improved data selection for domain adaptation in ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7018–7022, doi: 10.1109/ICASSP39728.2021.9413869.

[49] X. Sun, B. Wang, S. Liu, T. Lu, X. Shan, and Q. Yang, "LMC-SMCA: A new active learning method in ASR," *IEEE Access*, vol. 9, pp. 37011–37021, 2021, doi: 10.1109/ACCESS.2021.3062157.

[50] F. Mezzoudj, D. Langlois, D. Jouvet, and A. Benyettou, "Textual data selection for language modelling in the scope of automatic speech recognition," *Proc. Comput. Sci.*, vol. 128, pp. 55–64, Jan. 2018, doi: 10.1016/j.procs.2018.03.008.

[51] T. Nadungodage, R. Weerasinghe, and M. Niranjan, "Efficient use of training data for sinhala speech recognition using active learning," in *Proc. Int. Conf. Adv. ICT Emerg. Regions (ICTer)*, Dec. 2013, pp. 149–153, doi: 10.1109/ICTER.2013.6761170.

[52] Y. Khassanov, T. Y. Chong, B. Bigot, and E. S. Chng, "Unsupervised language model adaptation by data selection for speech recognition," in *Intelligent Information and Database Systems*. Cham, Switzerland: Springer, 2017, pp. 508–517.

[53] K. Kirchhoff, J. Bilmes, K. Wei, Y. Liu, A. Mandal, and C. Bartels, "A Submodularity Framework for Data Subset Selection," Air Force Res. Lab., Ohio, USA, Tech. Rep. AFRL-RH-WP-TR-2013-0108, 2013.

[54] E. Gouvêa and M. H. Davel, "Kullback–Leibler divergence-based ASR training data selection," in *Proc. Interspeech*, Aug. 2011, pp. 2297–2300.

[55] C. Pelaez-Moreno, Q. Zhu, B. Y. Chen, and N. Morgan, "Automatic data selection for MLP-based feature extraction for ASR," in *Proc. Interspeech*, Sep. 2005, pp. 229–232.

[56] M. Doulaty and T. Hain, "Latent Dirichlet allocation based acoustic data selection for automatic speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 3228–3232, doi: 10.21437/interspeech.2019-1797.

[57] J. Qi, X. Liu, S. Kamijo, and J. Tejedor, "Distributed submodular maximization for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2501–2505, doi: 10.1109/ICASSP.2018.8462240.

[58] C. Wu, R. W. M. Ng, O. S. Torralba, and T. Hain, "Analysing acoustic model changes for active learning in automatic speech recognition," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2017, pp. 1–5, doi: 10.1109/IWSSIP.2017.7965609.

[59] Y.-F. Cheng, L.-W. Chen, H.-S. Lee, and H.-M. Wang, "Exploring the impact of data quantity on ASR in extremely low-resource languages," 2024, *arXiv:2409.08872*.

[60] Y. Xi, W. Ding, K. Yu, and J. Lai, "Semi-supervised learning for code-switching ASR with large language model filter," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2024, pp. 263–270, doi: 10.1109/slt61566.2024.10832173.

[61] N. Ravi and V. Arora, "TeLeS: Temporal lexeme similarity score to estimate confidence in end-to-end ASR," *Proc. IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4399–4408, Jan. 2024, doi: 10.1109/TASLP.2024.3416687.

[62] L. Wang, "Revisit micro-batch clipping: Adaptive data pruning via gradient manipulation," 2024, *arXiv:2408.16204*.

[63] Z. Zheng, Z. Ma, Y. Wang, and X. Chen, "Unsupervised active learning: Optimizing labeling cost-effectiveness for automatic speech recognition," in *Proc. INTERSPEECH*, Aug. 2023, pp. 3307–3311, doi: 10.21437/interspeech.2023-614.

[64] I.-T. Hsieh, C.-H. Wu, and S.-W. Tsa, "Data selection based on phoneme affinity matrix for electrolarynx speech recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Oct. 2023, pp. 2196–2202, doi: 10.1109/apsipaasc58517.2023.10317555.

[65] M. K. Baskar, A. Rosenberg, B. Ramabhadran, Y. Zhang, and P. Moreno, "Ask2Mask: Guided data selection for masked speech modeling," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1357–1366, Oct. 2022, doi: 10.1109/JSTSP.2022.3186162.

[66] C. Park, R. Ahmad, and T. Hain, "Unsupervised data selection for speech recognition with contrastive loss ratios," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8587–8591, doi: 10.1109/ICASSP43922.2022.9747390.

[67] A. Afshan, K. Kumar, and J. Wu, "Sequence-level confidence classifier for ASR utterance accuracy and application to acoustic models," in *Proc. Interspeech*, Aug. 2021, pp. 4084–4088, doi: 10.21437/interspeech.2021-1666.

[68] W. Zheng, A. Xiao, G. Keren, D. Le, F. Zhang, C. Fuegen, O. Kalinli, Y. Saraf, and A. Mohamed, "Scaling ASR improves zero and few shot learning," in *Proc. Interspeech*, Sep. 2022, pp. 5135–5139, doi: 10.21437/interspeech.2022-11023.

[69] R. Gretter, M. Matassoni, and D. Falavigna, "Seed words based data selection for language model adaptation," 2021, *arXiv:2107.09433*.
[70] J. Bang, H. Kim, Y. Yoo, and J.-W. Ha, "Boosting active learning for speech recognition with noisy pseudo-labeled samples," 2020, *arXiv:2006.11021*.
[71] P. Swarup, D. Chakrabarty, A. Sapru, H. Tulsiani, H. Arsikere, and S. Garimella, "Knowledge distillation and data selection for semi-supervised learning in CTC acoustic models," 2020, *arXiv:2008.03923*.
[72] T. Drugman, J. Pylkkonen, and R. Kneser, "Active and semi-supervised learning in ASR: Benefits on the acoustic and language models," 2019, *arXiv:1903.02852*.
[73] J. Fu and K. Ru, "A dropout-based single model committee approach for active learning in ASR," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 16–22.
[74] C. Ni and B. Ma, "Investigate automatic speech recognition and keyword search for very low-resource language," in *Proc. IEEE 2nd Int. Conf. Signal Image Process. (ICSIP)*, Aug. 2017, pp. 336–340.
[75] M. Doulaty, O. Saz, and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proc. Interspeech*, Sep. 2015, pp. 2897–2901, doi: 10.21437/interspeech.2015-609.
[76] X. Pan, Y. Cao, and Y. Lu, "Active learning and semi-supervised learning in Tibetan language speech recognition," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, vol. 1, Oct. 2010, pp. 369–372, doi: 10.1109/AICI.2010.84.
[77] S. Li, Y. Akita, and T. Kawahara, "Automatic lecture transcription based on discriminative data selection for lightly supervised acoustic model training," *IEICE Trans. Inf. Syst.*, vol. E98.D, no. 8, pp. 1545–1552, 2015, doi: 10.1587/transinf.2015edp7047.
[78] A. Mittal, D. Sivasubramanian, R. Iyer, P. Jyothi, and G. Ramakrishnan, "Partitioned gradient matching-based data subset selection for compute-efficient robust ASR training," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2022, pp. 5999–6010, doi: 10.18653/v1/2022.findings-emnlp.443.
[79] E. Yilmaz, J. F. Gemmeke, and H. Van hamme, "Data selection for noise robust exemplar matching," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5980–5984, doi: 10.1109/ICASSP.2016.7472825.
[80] R. Petcu and S. Maji, "Efficient data selection employing semantic similarity-based graph structures for model training," 2024, *arXiv:2402.14888*.
[81] Q. Liu, W.-Q. Zhang, J. Liu, and Y. Liu, "A general procedure for improving language models in low-resource speech recognition," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2019, pp. 428–433, doi: 10.1109/ialp48816.2019.9037726.
[82] U. Nallasamy, F. Metze, and T. Schultz, "Active learning for accent adaptation in automatic speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2012, pp. 360–365.
[83] C. Ni, L. Wang, H. Liu, C.-C. Leung, L. Lu, and B. Ma, "Submodular data selection with acoustic and phonetic features for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4629–4633.
[84] S. Li, Y. Akita, and T. Kawahara, "Semi-supervised acoustic model training by discriminative data selection from multiple ASR systems' hypotheses," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1524–1534, Sep. 2016.
[85] Z. Lu, Y. Wang, Y. Zhang, W. Han, Z. Chen, and P. Haghani, "Unsupervised data selection via discrete speech representation for ASR," 2022, *arXiv:2204.01981*.
[86] C. Gao, G. Cheng, P. Zhang, and Y. Yan, "Speech corpora divergence based unsupervised data selection for ASR," 2023, *arXiv:2302.13222*.
[87] M. Baali, T. Hayashi, H. Mubarak, S. Maiti, S. Watanabe, W. El-Hajj, and A. Ali, "Unsupervised data selection for TTS: Using Arabic broadcast news as a case study," 2023, *arXiv:2301.09099*.
[88] A. W. Black and K. Lenzo, "Optimal utterance selection for unit selection speech synthesis databases," *Int. J. Speech Technol.*, vol. 6, no. 4, pp. 357–363, Oct. 2003.
[89] E. Tesfaye Biru, Y. Tofik Mohammed, D. Tofu, E. Cooper, and J. Hirschberg, "Subset selection, adaptation, gemination and prosody prediction for amharic text-to-speech synthesis," in *Proc. 10th ISCA Workshop Speech Synth. (SSW)*, Sep. 2019, pp. 205–210.
[90] F.-Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, and I. Ouyang, "Data selection for improving naturalness of TTS voices trained on small found corpuses," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 319–324.
[91] A. Chalamandaris, P. Tsiakoulis, S. Raptis, and S. Karabetsos, "Design of an efficient corpus for high-quality unit selection TTS for Bulgarian," Athena Res. Centre, Athens, Greece, Tech. Rep., 2009. [Online]. Available: http://speech.ilsp.gr/synthesis/publications/ssgpub_ltc_2009.pdf
[92] Z. Zhao and X. Ma, "Active learning for prediction of prosodic word boundaries in Chinese TTS using maximum entropy Markov model," *J. Softw.*, vol. 8, no. 12, pp. 3222–3228, Dec. 2013.
[93] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[94] J.-W. Jung, W. Zhang, S. Maiti, Y. Wu, X. Wang, J.-H. Kim, Y. Matsunaga, S. Um, J. Tian, H.-J. Shim, N. Evans, J. S. Chung, S. Takamichi, and S. Watanabe, "Text-to-speech synthesis in the wild," 2024, *arXiv:2409.08711*.
[95] N. Braunschweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality," in *Proc. Interspeech*, Aug. 2011, pp. 1821–1824.
[96] E. Cooper, A. Chang, Y. Levitan, and J. Hirschberg, "Data selection and adaptation for naturalness in HMM-based speech synthesis," in *Proc. Interspeech*, Sep. 2016, pp. 357–361.
[97] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, "Diversity-based core-set selection for text-to-speech with linguistic and acoustic features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 12351–12355.
[98] A. R. Mandeel, M. S. Al-Radhi, and T. G. Csapó, "Modeling irregular voice in end-to-end speech synthesis via speaker adaptation," in *Proc. Int. Conf. Speech Technol. Human-Computer Dialogue (SpeD)*, Oct. 2023, pp. 170–175.
[99] A. González-Docasal and A. Álvarez, "Enhancing voice cloning quality through data selection and alignment-based metrics," *Appl. Sci.*, vol. 13, no. 14, p. 8049, Jul. 2023.
[100] A. Singh, A. Nagireddi, J. Bandekar, S. Badiger, S. Udupa, P. K. Ghosh, H. A. Murthy, H. Zen, P. Kumar, K. Kant, A. Bole, B. C. Singh, K. Tokuda, M. Hasegawa-Johnson, and P. Olbrich, "Lightweight, multi-speaker, multi-lingual indic text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–2.
[101] T. Godambe, S. K. Rallabandi, S. V. Gangashetty, A. Alkhairy, and A. Jafri, "Developing a unit selection voice given audio without corresponding text," *EURASIP J. Audio, Speech, Music Process.*, vol. 2016, pp. 1–11, Mar. 2016.
[102] K.-Z. Lee, E. Cooper, and J. Hirschberg, "A comparison of speaker-based and utterance-based data selection for text-to-speech synthesis," in *Proc. Interspeech*, Sep. 2018, pp. 2873–2877.
[103] Z. Yu, K. Wang, Y. Zu, D. Yue, and G. Chen, "Data pruning approach to unit selection for inventory generation of concatenative embeddable Chinese TTS systems," in *Proc. Interspeech*, Oct. 2004, pp. 1177–1180.
[104] A. H. Azeemi, I. A. Qazi, and A. A. Raza, "Dataset pruning for resource-constrained spoofed audio detection," in *Proc. Interspeech*, Sep. 2022, pp. 416–420.
[105] A. H. Azeemi, I. A. Qazi, and A. A. Raza, "Self-supervised dataset pruning for efficient training in audio anti-spoofing," in *Proc. INTERSPEECH*, Aug. 2023, pp. 2773–2777.
[106] H. Dhamyal, A. Ali, I. A. Qazi, and A. A. Raza, "Fake audio detection in resource-constrained settings using microfeatures," in *Proc. Interspeech*, Aug. 2021, pp. 4149–4153.
[107] J. C. S. Valera, A. M. P. Herreros, and Á. M. G. García, "Data augmentation techniques for physical access in voice anti-spoofing," in *Proc. Iber Speech*, 2024, pp. 1–5, doi: 10.21437/IberSPEECH.2024-1.
[108] J. A. J. Garizao, "Implementation of active data selection algorithms for data choosing in ASV systems," Universidad de los Andes, Colombia, Tech. Rep., 2023. [Online]. Available: https://repositorio.uniandes.edu.co/bitstreams/f93beba8-a0cc-497c-8179-c6bbe3ecc332/download
[109] X. Wang and J. Yamagishi, "Investigating active-learning-based training data selection for speech spoofing countermeasure," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2023, pp. 585–592.
[110] C. Hanilçi, "Data selection for i-vector based automatic speaker verification anti-spoofing," *Digit. Signal Process.*, vol. 72, pp. 171–180, Jan. 2018, doi: 10.1016/j.dsp.2017.10.010.
[111] M. Lindsey, N. R. Robinson, F. Kubala, and R. M. Stern, "Reducing the cost of spoof detection labeling using mixed-strategy active learning and pretrained models," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2023, pp. 1–7, doi: 10.1109/asru57964.2023.10389626.

[112] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2507–2522, 2023.
[113] S. Munir, W. Sajjad, M. Raza, E. Abbas, A. H. Azeemi, I. A. Qazi, and A. A. Raza, "Deepfake defense: Constructing and evaluating a specialized Urdu deepfake audio dataset," in *Proc. Findings Assoc. Comput. Linguistics ACL*, 2024, pp. 14470–14480.
[114] X. Qi, H. Gu, J. Yi, J. Tao, Y. Ren, J. He, and S. Zeng, "MADD: A multilingual multi-speaker audio deepfake detection dataset," in *Proc. IEEE 14th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Nov. 2024, pp. 466–470.
[115] H. Lei, "Towards structured approaches to arbitrary data selection and performance prediction for speaker recognition," in *Proc. Int. Conf. Biometrics*. Cham, Switzerland: Springer, Jan. 2009, pp. 513–522.
[116] R. Abdullah, V. Vijean, H. Muthusamy, F. N. Che Kassim, Z. Abdullah, M. N. Md Noor, and J. A. R. Rawi, "Wavelet-based parametric feature subset selection for speaker and accent recognition using genetic algorithm," *J. Telecommun., Electron. Comput. Eng. (JTEC)*, vol. 15, no. 1, pp. 29–37, Mar. 2023.
[117] C.-L. Huang and B. Ma, "Maximum entropy based data selection for speaker recognition," in *Proc. Interspeech*, Aug. 2011, pp. 2713–2716.
[118] D. Pullella, M. Kuhne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4833–4836.
[119] H.-W. Sun, B. Ma, and H.-Z. Li, "An efficient feature selection method for speaker recognition," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, Dec. 2008, pp. 181–184.
[120] D. Nath and S. Kr. Kalita, "Feature selection method for speaker recognition using neural network," *Int. J. Comput. Appl.*, vol. 101, no. 3, pp. 38–44, Sep. 2014.
[121] A. Harrag, D. Saigaa, K. Boukharouba, M. Drif, and A. Bouchelaghem, "GA-based feature subset selection: Application to Arabic speaker recognition system," in *Proc. 11th Int. Conf. Hybrid Intell. Syst. (HIS)*, Dec. 2011, pp. 383–387.
[122] S. H. Shum, N. Dehak, and J. R. Glass, "Limited labels for unlimited data: Active learning for speaker recognition," in *Proc. Interspeech*, Sep. 2014, pp. 383–387, doi: 10.21437/interspeech.2014-89.
[123] A. Harrag, "Nature-inspired feature subset selection application to Arabic speaker recognition system," *Int. J. Speech Technol.*, vol. 18, no. 2, pp. 245–255, Dec. 2014, doi: 10.1007/s10772-014-9264-2.
[124] S. Biswas, J. Rohdin, and K. Shinoda, "Autonomous selection of i-vectors for PLDA modelling in speaker verification," *Speech Commun.*, vol. 72, pp. 32–46, Sep. 2015, doi: 10.1016/j.specom.2015.05.001.
[125] C.-L. Huang, J.-C. Wang, and B. Ma, "Ensemble based speaker recognition using unsupervised data selection," *APSIPA Trans. Signal Inf. Process.*, vol. 5, no. 1, 2016, doi: 10.1017/atsip.2016.10. [Online]. Available: https://www.nowpublishers.com/article/Details/SIP-062
[126] H. Lei and E. Lopez-Gonzalo, "Importance of nasality measures for speaker recognition data selection and performance prediction," in *Proc. Interspeech*, Sep. 2009, pp. 888–891, doi: 10.21437/interspeech.2009-268.
[127] S. Karlos, K. Kaleris, N. Fazakis, V. G. Kanas, and S. Kotsiantis, "Optimized active learning strategy for audiovisual speaker recognition," in *Speech and Computer*. Cham, Switzerland: Springer, 2018, pp. 281–290.
[128] G. Sarkar and G. Saha, "Spectral entropy and spectral shape based pre-quantization for real time speaker identification system," *Int. J. Speech Technol.*, vol. 13, no. 4, pp. 189–199, Oct. 2010, doi: 10.1007/s10772-010-9079-8.
[129] C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, "Ensemble classifiers using unsupervised data selection for speaker recognition," in *Proc. Interspeech*, Sep. 2012, pp. 2666–2669, doi: 10.21437/interspeech.2012-348.
[130] H. Lei and N. Mirghafori, "Data selection with kurtosis and nasality features for speaker recognition," in *Proc. Interspeech*, Aug. 2011, pp. 2753–2756, doi: 10.21437/interspeech.2011-155.
[131] C.-L. Huang and H. Li, "UBM data selection for effective speaker modeling," in *Proc. 7th Int. Symp. Chin. Spoken Lang. Process.*, Nov. 2010, pp. 162–165, doi: 10.1109/ISCSLP.2010.5684493.
[132] J. H. L. Hansen, J.-W. Suh, P. Angkititrakul, and Y. Lei, "Effective background data selection for SVM-based speaker recognition with unseen test environments: More is not always better," *Int. J. Speech Technol.*, vol. 17, no. 3, pp. 211–221, Jan. 2014, doi: 10.1007/s10772-013-9219-z.
[133] A. Harrag, D. Saigaa, A. Bouchelaghem, M. Drif, S. Zeghlache, and N. Harrag, "How to reduce dimension while improving performance," in *Hybrid Artificial Intelligent Systems*. Berlin, Germany: Springer, 2012, pp. 497–508.
[134] K.-P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Jul. 1988, pp. 595–598, doi: 10.1109/ICASSP.1988.196655.
[135] D. Le and E. M. Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 146–152.
[136] W. Han, H. Li, H. Ruan, L. Ma, J. Sun, and B. Schuller, "Active learning for dimensional speech emotion recognition," in *Proc. Interspeech*, Aug. 2013, pp. 2841–2845.
[137] Z. Zhao and X. Ma, "Active learning for speech emotion recognition using conditional random fields," in *Proc. 14th ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput.*, Jul. 2013, pp. 127–131.
[138] M. Abdelwahab and C. Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5160–5164.
[139] D. Li, Y. Wang, K. Funakoshi, and M. Okumura, "After: Active learning based fine-tuning framework for speech emotion recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2023, pp. 1–8.
[140] F. Ren, Z. Liu, and X. Kang, "An efficient framework for constructing speech emotion corpus based on integrated active learning strategies," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1929–1940, Oct. 2022.
[141] H.-C. Lin, Y.-C. Lin, H.-C. Chou, and H.-Y. Lee, "Improving speech emotion recognition in under-resourced languages via speech-to-speech translation with bootstrapping data selection," 2024, *arXiv:2409.10985*.
[142] X. Li and M. Akagi, "Maximal information coefficient and predominant correlation-based feature selection toward a three-layer model for speech emotion recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 1428–1434.
[143] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015, doi: 10.1109/TASLP.2014.2375558.
[144] I. Kansizoglou, L. Bampis, and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 756–768, Apr. 2022, doi: 10.1109/TAFFC.2019.2961089.
[145] S. Kuchibhotla, H. D. Vankayalapati, and K. R. Anne, "An optimal two stage feature selection for speech emotion recognition using acoustic features," *Int. J. Speech Technol.*, vol. 19, no. 4, pp. 657–667, Aug. 2016, doi: 10.1007/s10772-016-9358-0.
[146] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc. Interspeech*, Sep. 2012, pp. 362–365, doi: 10.21437/interspeech.2012-117.
[147] J.-B. Kim and J.-S. Park, "Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition," *Eng. Appl. Artif. Intell.*, vol. 52, pp. 126–134, Jun. 2016, doi: 10.1016/j.engappai.2016.02.018.
[148] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 275–278, doi: 10.1145/2818346.2820774.
[149] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5000–5004, doi: 10.1109/ICASSP.2017.7953108.
[150] A. Álvarez, B. Sierra, A. Arruti, J.-M. López-Gil, and N. Garay-Vitoria, "Classifier subset selection for the stacked generalization method applied to emotion recognition in speech," *Sensors*, vol. 16, no. 1, p. 21, Dec. 2015, doi: 10.3390/s16010021.
[151] A. Arruti, I. Cearreta, A. Álvarez, E. Lazkano, and B. Sierra, "Feature selection for speech emotion recognition in Spanish and basque: On the use of machine learning to improve human–computer interaction," *PLoS ONE*, vol. 9, no. 10, Oct. 2014, Art. no. e108975, doi: 10.1371/journal.pone.0108975.

[152] Z. Zhang, J. Deng, E. Marchi, and B. Schuller, "Active learning by label uncertainty for acoustic emotion recognition," in *Proc. Interspeech*, Aug. 2013, pp. 2856–2860, doi: 10.21437/interspeech.2013-250.
[153] A. Álvarez, I. Cearreta, J. M. López, A. Arruti, E. Lazkano, B. Sierra, and N. Garay-Vitoria, "Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken Spanish and standard Basque language," in *Proc. Int. Conf. Text, Speech Dialogue*. Cham, Switzerland: Springer, Jan. 2006, pp. 565–572.
[154] D. Su and P. Fung, "Personalized music emotion classification via active learning," in *Proc. 2nd Int. ACM workshop Music Inf. Retr. with user-centered multimodal strategies*, Nov. 2012, pp. 57–62, doi: 10.1145/2390848.2390864.
[155] J. Esparza, S. Scherer, and F. Schwenker, "Studying self- and active-training methods for multi-feature set emotion recognition," in *Partially Supervised Learning*. Berlin, Germany: Springer, 2012, pp. 19–31.
[156] S. Meudt and F. Schwenker, "On instance selection in audio based emotion recognition," in *Artificial Neural Networks in Pattern Recognition*. Berlin, Germany: Springer, 2012, pp. 186–192.
[157] S. Hantke, A. Abstreiter, N. Cummins, and B. Schuller, "Trustability-based dynamic active learning for crowdsourced labelling of emotional audio data," *IEEE Access*, vol. 6, pp. 42142–42155, 2018, doi: 10.1109/ACCESS.2018.2858931.
[158] X. Jiang, L. Meng, and D. Wu, "Multi-task active learning for simultaneous emotion classification and regression," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 1947–1952, doi: 10.1109/SMC52423.2021.9659052.
[159] F. Haider, S. Pollak, P. Albert, and S. Luz, "Extracting audio-visual features for emotion recognition through active feature selection," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5, doi: 10.1109/GlobalSIP45357.2019.8969360.
[160] E. Bozkurt, E. Erzin, Ç. E. Erdem, and A. T. Erdem, "RANSAC-based training data selection on spectral features for emotion recognition from spontaneous speech," in *Proc. Int. Conf. Anal. Verbal Nonverbal Commun. Enactment*. Cham, Switzerland: Springer, Jan. 2011, pp. 36–47.
[161] X. Yu, "Adaptability of simple classifier and active learning in music emotion recognition," in *Proc. 4th Int. Conf. Electron., Commun. Control Eng.*, Apr. 2021, pp. 13–19, doi: 10.1145/3462676.3462679.
[162] S. A. Moreno-Acevedo, J. C. Vasquez-Correa, J. M. Martín-Doñas, and A. Álvarez, "Stream-based active learning for speech emotion recognition via hybrid data selection and continuous learning," in *Text, Speech, and Dialogue*. Cham, Switzerland: Springer Nature Switzerland, 2024, pp. 105–117.
[163] A. Álvarez, I. Cearreta, J. M. López, A. Arruti, E. Lazkano, B. Sierra, and N. Garay-Vitoria, "A comparison using different speech parameters in the automatic emotion recognition using feature subset selection based on evolutionary algorithms," in *Proc. Int. Conf. text, speech dialogue*, Jan. 2007, pp. 423–430.
[164] A. Álvarez, I. Cearreta, J. M. López, A. Arruti, E. Lazkano, B. Sierra, and N. Garay-Vitoria, "Application of feature subset selection based on evolutionary algorithms for automatic emotion recognition in speech," in *Proc. Int. Conf. Nonlinear Speech Process.*, Jan. 2007, pp. 273–281.
[165] G. Dahy, A. E. Hassanien, and S. H. Basha, "Hybrid intelligent model for speech emotion recognition using active learning and residual network," in *Proc. Int. Conf. Adv. Intell. Syst. Informat.*, Jan. 2025, pp. 488–497.
[166] Y. Xu, X. Jiang, and D. Wu, "Cross-task inconsistency based active learning (CTIAL) for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1659–1668, Jul. 2024.
[167] D. Li, Y. Zhang, Y. Wang, K. Funakoshi, and M. Okumura, "Active learning with task adaptation pre-training for speech emotion recognition," *J. Natural Lang. Process.*, vol. 31, no. 3, pp. 825–867, 2024.
[168] E. Vaaras, M. Airaksinen, and O. Räsänen, "Analysis of self-supervised learning and dimensionality reduction methods in clustering-based active learning for speech emotion recognition," 2022, *arXiv:2206.10188*.
[169] C. Yu and J. H. L. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2188–2198, Nov. 2017.
[170] I. López-Espejo, Z.-H. Tan, J. H. Hansen, and J. Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, vol. 10, pp. 4169–4199, 2022.
[171] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

ABDUL HAMEED AZEEMI received the B.S. degree in computer science from FAST NUCES, Lahore, Pakistan, and the M.S. degree in data science from Information Technology University, Lahore. He is currently pursuing the Ph.D. (CS) degree with LUMS, Pakistan. His work has been published in premier NLP and Speech conferences, including ACL, EMNLP, COLING, and Interspeech. His research interests include data-efficient machine learning, speech processing, and natural language processing.

IHSAN AYYUB QAZI received the B.Sc. degrees (Hons.) in computer science and mathematics from LUMS, Pakistan, in 2005, and the Ph.D. degree in computer science from the University of Pittsburgh, PA, USA, in 2010. He is a tenured Full Professor with the Department of Computer Science, LUMS. He was a Visiting Research Scientist at the University of California at Berkeley, Berkeley, USA, in 2017. Previously, he has held positions at BBN Technologies, Cambridge, USA; and the Centre for Advanced Internet Architectures, Melbourne, Australia. His research interests are in networked systems and span cloud computing, ICT for development, mobile/wireless networks, and online privacy. His work has appeared in premier networking conferences, such as ACM SIGCOMM, and journals, including IEEE/ACM TRANSACTIONS ON NETWORKING. His work was selected as the Best ACM SIGCOMM CCR Paper in 2018. He was a recipient of the Google Faculty Research Award, the Facebook Research Award, the Sheth International Young Alumni Achievement Award 2021 (University of Pittsburgh), the LUMS Vice Chancellor's Alumni Achievement Award, and the Andrew Mellon Fellowship. His research has been sponsored by U.S. National Science Foundation (NSF), World Bank, Australian Research Council (ARC), Facebook, Google, HEC, and Ignite.

AGHA ALI RAZA received the Ph.D. degree in computer science from Carnegie Mellon University, Pittsburgh, USA, with a specialization in language technologies. He is a tenured Associate Professor with the Department of Computer Science, LUMS, Pakistan. He is a Fulbright Scholar. His research focuses on speech and natural language processing, artificial intelligence (AI) and machine learning (ML), and technologies for emerging markets. His innovative products, software, and services have served more than 300 000 low-literate and non-tech-savvy people in Pakistan, India, and West Africa, leading to a significant social, health, and economic impact. He has been awarded the Highly Competitive Grants by UNICEF, GIZ, Meta, the Gates Foundation, NIH, the National Academies of Sciences (NAKFI), Google Inc., and HEC. His projects have been featured in prestigious national and international media, including BBC, NPR, MIT Tech Review International, Express Tribune, Dawn News, Geo News, Aaj TV, Neo TV, Times of India, IEEE Signal Processing Society, Pittsburgh Post-Gazette, and Voice of America.