Towards Efficient Universal Audio Analysis: A Low-Complexity Model via Synergistic Multi-Task Learning

Maxim K. Surkov
ITMO University
Saint Petersburg, Russia
surkovmax007@mail.ru

***Abstract*—Audio analysis is a cornerstone of modern human-computer interaction, powering applications in smart devices such as phones, watches, and speakers. Key tasks enabling this interaction include voice activity detection, speech command recognition, and age and gender identification. While large-scale models can seamlessly integrate these tasks without significant performance degradation, their computational cost prohibits deployment on resource-constrained devices. In contrast, compact models suitable for embedded systems face a significant challenge: their limited parameter budget makes integrating multiple tasks without destructive interference difficult, and the optimal weight-sharing strategy is highly dependent on the specific task combination. This paper presents an analysis of weight-sharing architectures for multi-task learning on audio data. We systematically investigate the synergies and conflicts between tasks, evaluating pairwise and higher-order combinations. Based on our findings, we propose a novel low-complexity model that simultaneously executes four core tasks—voice activity detection, speech command recognition, age and gender identification—with fewer than 57,000 parameters. This represents a 53% reduction in model size compared to a naive ensemble of single-task models. Our work not only achieves state-of-the-art performance on each individual task but also reveals that while most tasks exhibit positive synergy, the introduction of a more complex task, such as speaker diarization, can lead to performance degradation in larger task sets, highlighting the importance of careful task selection and architectural design for stable multi-task learning.**

I. Introduction

The proliferation of intelligent devices—including smartphones, smart speakers, and wearables—has driven increasing demand for efficient, on-device audio analysis systems. Such systems are essential for enabling natural human–computer interaction (HCI) through a suite of fundamental audio understanding tasks, including voice activity detection (VAD), speech command recognition (SCR), and speaker biometrics such as age and gender classification. Integrating these capabilities allows devices to build richer contextual awareness of users and their environments, supporting more personalized and responsive interactions. For example, a device may adapt its acoustic model based on inferred speaker characteristics or use VAD to identify segments where more resource-intensive server-side processing may be applied, thereby reducing overall latency.

State-of-the-art approaches to these tasks typically employ deep learning architectures, often combining multi-layer convolutional neural networks (CNNs) with recurrent layers or attention mechanisms for temporal modeling [1]–[6]. A common pipeline involves first converting raw audio waveforms into log-mel spectrograms via the Short-Time Fourier Transform (STFT). A CNN then processes this feature sequence into encoded representations, which are subsequently transformed into final predictions using either linear projections (for global classification) or recurrent neural networks (for frame-wise predictions).

Notable standalone models include the SILERO framework [7] for VAD, which achieves ROC-AUC scores 94% on benchmarks such as AI-SHELL-4 [8] and ALI-MEETINGS [9] utilizing 260,000 parameters. For speaker biometrics, recent work [4] reports error rates of 3.6% for age classification and 30% for gender classification on the Mozilla Common Voice dataset [10] using the CNN model with 30,000 trainable weights. In speech command recognition, top-performing models [11] reach accuracies of 95.3% on the Google Speech Commands V2 dataset [12]. While performant, deploying separate models for each task incurs a linear increase in memory usage with the number of tasks, rendering this approach impractical for resource-constrained devices.

Multi-task learning (MTL) [13] presents a promising alternative, though it is predominantly applied in large-scale models (e.g., Qwen-Audio [14]) with billions of parameters, unsuitable for on-device inference [15]–[23]. Naively applying MTL to low-complexity models introduces two significant challenges: (1) determining optimal weight-sharing strategies across tasks to maintain performance, and (2) selecting compatible tasks that exhibit positive synergies without destructive interference.

In this work, we present a study of architectural and task-compositional factors in low-complexity multi-task audio models. We systematically evaluate weight-sharing configurations and task subsets to identify synergistic combinations and avoid detrimental interactions. Furthermore, we explore the impact of incorporating more complex tasks, such as speaker diarization, and demonstrate that improper task inclusion can significantly degrade performance. Based on our analysis, we propose efficient multi-task architectures that simultaneously address multiple audio tasks with minimal parameter overhead,

achieving state-of-the-art accuracy while significantly reducing model size compared to naive ensemble of single-task baselines.

Our key contributions are:

1) A systematic analysis of weight-sharing strategies in low-complexity multi-task audio models, including an evaluation of pairwise and higher-order task synergies.
2) A parameter-efficient multi-task architecture that simultaneously supports four core tasks—VAD, SCR, age, and gender classification—with under 57K parameters, achieving state-of-the-art accuracy while reducing model size by 53%.
3) A key insights revealing that while many tasks benefit from joint learning, introducing complex tasks (e.g., diarization) can impair performance in larger task sets, highlighting the importance of careful task selection and model design for on-device MTL.

This paper is organized as follows. In Section II, we detail our proposed method, including the baseline single-task architectures and the systematic multi-task learning framework with three distinct weight-sharing strategies. In Section III we describe the experimental setup, specifying the datasets, training procedures, evaluation metrics, and model configurations used for our analysis. In Section IV, we present and analyze the results, examining the impact of weight-sharing ratios, quantifying pairwise and higher-order task synergies, and evaluating the disruptive effect of integrating complex tasks. Finally, in Section V, we discuss the limitations of our work and suggest future research directions, and in Section VI, we conclude by summarizing our key findings and the successful development of a parameter-efficient unified audio model.

II. Method

A. Conventional single-task classification

This section outlines the conventional methodologies for the core tasks under investigation: VAD, SCR, and age and gender classification. Additionally, we incorporate the task of two-speaker diarization, defined as the process of determining the temporal segments during which each of two distinct speakers is active in an audio signal. Formally, our prediction model is defined by the operation $Y = D_{\theta_D}(E_{\theta_E}(X))$. Here, E denotes an encoder and D a decoder, with θ_E and θ_D representing their respective sets of trainable parameters. The input $X = (x_1, \ldots, x_N)$ is a sequence of log-mel spectrogram feature vectors, where each $x_i \in \mathbf{R}^D$ and N is the total number of time frames. The encoder produces a sequence of embeddings $E_{\theta_E}(X) = Q = (q_1, \ldots, q_N)$, which the decoder then maps to the final output $Y = D_{\theta_D}(Q)$. The total model complexity is consequently defined as the number of trainable parameters, given by $|\theta_E| + |\theta_D|$.

These five tasks can be categorized into two distinct groups based on their output structure: global classification and frame-wise classification. The first group, comprising SCR and age and gender classification, requires the model to produce a single label probability for the entire audio waveform, denoted as $Y \in \mathbf{R}^C$, where C is the total number of predefined classes. The second group consists of VAD and speaker diarization, which require a sequence of outputs. For VAD, the output is $Y \in \mathbf{R}^N$, representing the probability of speech presence for each time frame. For speaker diarization, the output is $Y \in \mathbf{R}^{N\times 2}$, indicating the probability of each of the two speakers being active at every time frame.

In this work, we employ a conventional multi-layer convolutional neural network as the encoder E, optionally augmented with an attention mechanism at its final stage. For global classification tasks, the decoder D consists of a single linear layer that projects the encoded embedding into the log-probabilities of predefined classes. In contrast, for frame-wise classification tasks, the decoder is implemented as a bidirectional Gated Recurrent Unit (GRU), followed by a linear projection layer to produce the final sequence of output predictions.

B. Conventional multi-task learning

Multi-task learning (MTL) is a well-established paradigm in machine learning that enables a single model to address multiple related tasks concurrently. This is typically achieved through a shared encoder network, which extracts a common latent representation from the input, coupled with a set of task-specific decoders. Formally, an MTL model comprises a shared encoder $E : X \to Q$ and a collection of t distinct decoders $D_1, \ldots, D_t$, where t denotes the number of tasks. Each decoder D_i transforms the shared representation Q into task-specific predictions Y_i. A principal advantage of this architecture is the significant improvement in computational and memory efficiency resulting from the reuse of the encoder's output across all tasks during inference.

The parameter efficiency of MTL can be quantified by comparison to a naive baseline employing t independent single-task models. This baseline requires t dedicated encoders $E_1, \ldots, E_t$ and t decoders $D_1, \ldots, D_t$, resulting in a total parameter count of $\sum_{i=1}^{t} |E_i| + \sum_{i=1}^{t} |D_i|$. In contrast, the MTL framework utilizes a single encoder E and t decoders, requiring only $|E| + \sum_{i=1}^{t} |D_i|$ parameters. Under the assumption of architecturally similar encoders ($|E_i| \approx |E| \ \forall i$), the parameter savings approximate $(t-1)|E|$. As the encoder generally constitutes the majority of a model's parameters, this reduction is substantial. Furthermore, the architecture is highly scalable; integrating an additional task introduces only the parameters of a new task-specific decoder D_{t+1}, indicating that the marginal cost of expanding the task set is low once the shared encoder is established.

However, applying MTL to low-complexity models introduces significant constraints. The strictly limited parameter budget can hinder the model's capacity to achieve competitive accuracy across all tasks simultaneously, a problem seldom encountered in large-scale models due to their abundant representational capacity. Consequently, the selection of a compatible task set and a deliberate weight-sharing strategy becomes critical. Careful architectural design is essential to mitigate destructive interference and leverage potential syner-

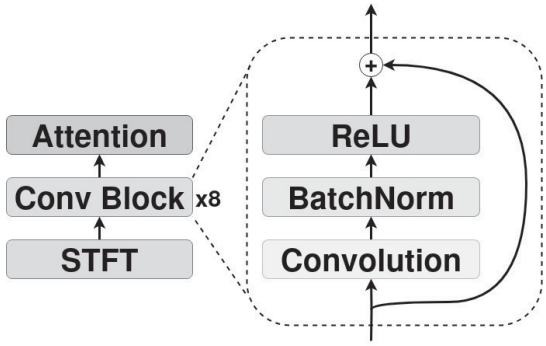
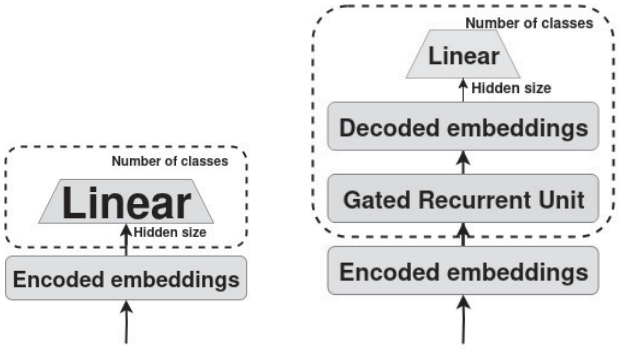


Fig. 1. Architecture of the multi-layer convolution neural network



(a) Global classification decoder (b) Frame-wise classification decoder

Fig. 2.Architecture of the classification decoders

TABLE I. A COMPARISON OF STATE-OF-THE-ART APPROACHES WITH OUR SINGLE-TASK MODELS

Task	Metric	Dataset	Baseline	Ours
VAD	ROC-AUC	AI-SHELL-4 [8]	94% [7]	93.3%
GC[1]	ER	MCV [10]	3.6% [4]	3.8%
AC[2]	ER	MCV [10]	30% [4]	32.8%
SCR	Accuracy	GSC v2 [12]	95.3% [11]	94.5%

[1] Gender classification
[2] Age group classification

gies between tasks to ensure robust performance across the entire multi-task system.

C. The proposed method

Based on the preceding analysis, two primary research questions emerge: (i) how are the tasks in the target set correlated, and (ii) how should parameters be shared between tasks exhibiting either strong or weak synergistic potential? To address these questions systematically, we propose the following experimental methodology.

First, we establish strong single-task baselines using a widely-adopted architectural template. Each individual model consist of an encoder comprising eight convolutional layers, each followed by batch normalization, ReLU activation function, and a residual connection. Optionally, an attention mechanism may be applied across the time axis atop the convolutional stack (see Fig. 1). For tasks requiring a global prediction (e.g., command recognition, age, or gender classification), the decoder consists of a single linear projection layer. For frame-wise prediction tasks (e.g., voice activity detection, diarization), the decoder is implemented as a bidirectional gated recurrent unit (GRU), followed by a linear layer to produce probabilities at each time step as shown in Fig. 2. This architecture first is trained and evaluated individually on each task to reproduce near state-of-the-art performance, thereby validating the baseline design (see Table I).

Subsequently, we conduct an exhaustive multi-task learning (MTL) analysis over all possible subsets of tasks. For each task combination, a multi-task model is constructed using a shared encoder and task-specific decoders. To evaluate the impact of weight-sharing strategy, each such model is trained and evaluated under three distinct sharing configurations (see Fig. 3):

1) Partial Sharing (7 Layers): The first seven convolutional layers are shared; the eighth layer and the optional attention mechanism are kept task-specific.
2) Full CNN Sharing (8 Layers): All eight convolutional layers are shared; the optional attention mechanism remains task-specific.
3) Complete Sharing (8 Layers + Attention): The entire encoder, including all eight convolutional layers and the attention mechanism, is shared across all tasks.

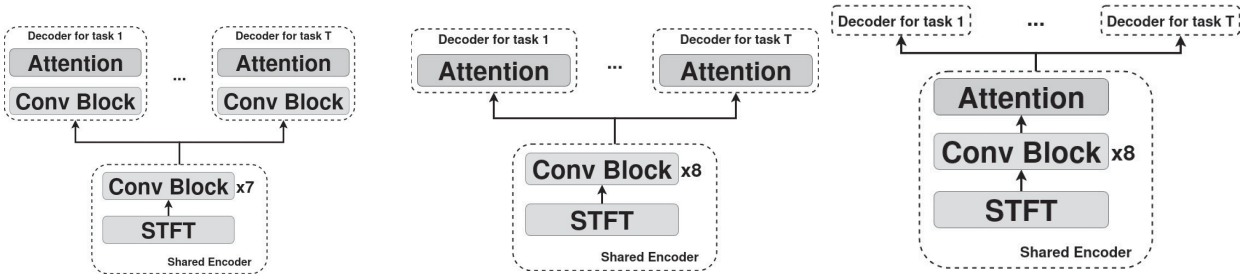
This structured ablation study enables a rigorous quantification of how the proportion of shared parameters influences final prediction accuracy across each individual task within every possible combination, thereby directly addressing our core research questions regarding task synergy and optimal sharing strategy.

III. EXPERIMENTS

A. Dataset

For the VAD task, model training and validation were conducted using the Mozilla Common Voice dataset [10], where human speech is present in approximately 75% of the annotated time segments. To ensure a rigorous comparison with state-of-the-art methods, evaluation was performed on the established benchmark datasets AI-SHELL-4 [8] and ALI-MEETINGS [9]. These benchmarks feature a higher speech density, with approximately 90% of segments containing speech. A strong correlation was observed between model performance on these two test sets. Therefore, for brevity and to simplify the analysis, we report detailed results only on the AI-SHELL-4 dataset in the subsequent sections.

The same Mozilla Common Voice [10] dataset was also used for the biometric tasks of age and gender classification. Samples with missing gender or age metadata were excluded during preprocessing. The final curated dataset contained approximately 500,000 samples of male speech and 200,000 samples of female speech.



(a) Partial Sharing (7 Layers) (b) Full CNN Sharing (8 Layers) (c) Complete Sharing (8 Layers + Attention)

Fig. 3. Multi-task learning weights sharing configurations

TABLE II. A SUMMARY OF THE DATASETS STATISTICS

Dataset	Number of samples	Duration in hours
AI-SHELL-4 [8]	3,065	12.7
ALI-MEETINGS [9]	10,379	43
MCV[1] [10] test	16,263	27
MCV [10] train	1,087,672	1718
GSC[2] v2 test [12]	11,005	3
GSC v2 train [12]	84,849	23

[1] Mozilla Common Voice
[2] Google Speech Commands

For age classification, only samples with defined age groups were retained, resulting in the following distribution: 300,000 samples from speakers in their twenties, 150,000 in their thirties, 111,000 in their forties, 75,000 from teenagers, 70,000 in their fifties, 61,000 in their sixties, 6,000 in their seventies, 1,000 in their eighties, and 178 samples from speakers in their nineties. Following the methodology of the baseline model [4], these samples were grouped into three broader age categories: under 30, 30 to 60, and over 60. The Mozilla Common Voice dataset was also used for testing these biometric tasks.

Finally, for the SCR task, the Google Speech Commands V2 [12] dataset was employed. This corpus contains 35 unique command classes, with each command represented by 1,000 to 3,000 samples in the training set and 150 to 450 samples in the test set.

A summary of the datasets statistics is provided in Table II. All audio samples are resampled to a 16 kHz sampling rate and subsequently transformed into log-mel spectrograms. These spectrograms are generated using 64-channel filter banks, computed over a 20 ms window with a 10 ms stride.

B. Training

A standardized training pipeline was employed for all experiments. For global classification tasks, the cross-entropy loss function was used, while binary cross-entropy was applied to frame-wise classification tasks. Optimization was performed using the AdamW optimizer with a weight decay of 0.01 and a batch size of 64.

The learning rate was scheduled using a cosine annealing strategy with an initial linear warmup phase. The warmup period comprised the first 5% of the total training epochs, during which the learning rate was linearly increased to a maximum value of 10^{-3}. For the remaining 95% of the training, the learning rate was decayed following a cosine annealing schedule. All models were trained for 100,000 epochs or until convergence was observed.

For MTL experiments, the overall loss was defined as a linear combination of the individual task losses with equal coefficients. Training batches were constructed by sampling uniformly across all tasks within the considered set. Model checkpoints were selected to minimize the worst-case relative performance degradation across all tasks, thereby ensuring balanced learning without significant compromise on any single objective.

The experimental results showed low variance, with standard deviations across multiple runs remaining within 1% of the reported metric values for all configurations. A single training run required approximately 10 hours of computation on an NVIDIA Tesla H100 GPU with 80GB of memory.

C. Evaluation metrics

Model performance was evaluated using standard task-specific metrics. For VAD, we employed the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). This metric evaluates the model's ability to distinguish between speech and non-speech segments across all classification thresholds, providing a robust single-figure measure of detection quality that is independent of the chosen operating point. For the biometric tasks of age and gender classification, performance was quantified using the Error Rate (ER), calculated as the proportion of incorrect predictions to total predictions. A lower ER indicates higher accuracy in classifying speaker demographics. SCR was assessed using classification accuracy, defined as the percentage of correctly identified voice commands within the test set. This provides a direct and intuitive measure of the model's practical utility for command-based interaction. In addition to task performance, we reported the overall model complexity in terms of the total number of trainable parameters. This metric is critical for determining the feasibility of deployment on resource-constrained devices and for comparing the architectural efficiency of different multi-task learning configurations.

TABLE III. MULTI-TASK LEARNING PERFORMANCE FOR ALL PAIRWISE TASK COMBINATIONS. THE UPPER-RIGHT PART REPORTS ABSOLUTE PERFORMANCE METRICS FOR MODELS TRAINED ON TWO TASKS. THE LOWER-LEFT PART QUANTIFIES THE RELATIVE PERFORMANCE REDUCTION COMPARED TO SINGLE-TASK BASELINES

Accuracy	**VAD**			**GC**[1]			**AC**[2]			**SCR**		
VAD		93.3		93.1/4	93/4.1	93.6/4.2	91.9/34	91.7/33.6	91.6/34	92.9/94	92.4/94.1	92.7/93.3
GC[1]	-0.3%	-0.33%	-0.4%		3.8		4.3/33.6	4.6/33.5	5.1/33.7	4.6/94.2	4.9/94.2	11.3/89.2
AC[2]	-1.7%	-1.76%	-1.85%	-1.2%	-1%	-1.4%		32.8		34/92.8	37.6/93.6	39.9/90.9
SCR	-0.6%	-1%	-1.3%	-0.8%	-1.2%	-7.8%	-2%	-7%	-10.5%		94.5	

Size	**VAD**	**GC**[1]	**AC**[2]	**SCR**
VAD	32K	47K(-23%)	47K(-23%)	47K(-23%)
GC[1]		29K	34K(-41%)	34K(-41%)
AC[2]			29K	34K(-41%)
SCR				30K

[1] Gender classification
[2] Age group classification

D. Configurations

We now provide a detailed description of our model architecture. The encoder network comprises eight sequential layers with 21 channels in each layer. Each layer consists of a two-dimensional convolutional operation, followed by batch normalization and a Rectified Linear Unit (ReLU) activation function. Residual connections are incorporated between layers that maintain an equivalent number of channels to facilitate gradient flow and stabilize the training process. An attention mechanism is applied over the temporal axis atop the encoder to enhance the integration of long-range contextual features.

The chosen architecture contains approximately 30,000 trainable parameters. This specific capacity was selected for several reasons. First, it aligns with the size of high-performing baseline models for biometric tasks [4]. Second, while state-of-the-art dedicated models for voice activity detection can be significantly larger (e.g., 260K parameters in [7]), our objective was to develop a compact multi-task architecture. Finally, through ablation studies, we determined that this parameter budget is sufficient for a single-task convolutional model to achieve competitive, near-state-of-the-art performance on the speech command recognition task, establishing it as a valid baseline for a efficient model.

Decoder architectures are task-dependent. For global classification tasks (e.g., age, gender, and speech command recognition), the decoder consists of a single linear projection layer that maps the hidden representation to the number of target classes. For frame-wise classification tasks (e.g., voice activity detection), the decoder is implemented as a multi-layer bidirectional Gated Recurrent Unit (GRU), which processes the temporal sequence of features to produce per-frame predictions.

In the MTL setup, the encoder is shared across tasks and uses the same hyperparameters as in the single-task models. The decoders remain task-specific and are identical to those used in the single-task experiments. For each subset of tasks, we investigate three distinct weight-sharing configurations:

1) Partial Sharing: The first seven convolutional layers are shared; the eighth layer and the attention mechanism are task-specific.
2) Full CNN Sharing: All eight convolutional layers are shared; the attention mechanism remains task-specific.
3) Complete Sharing: The entire encoder, including all convolutional layers and the attention mechanism, is shared across tasks.

To further demonstrate the critical importance of task selection in MTL, we introduce an additional complex task—two-speaker diarization—and analyze its impact on model performance when integrated into the existing task sets. This allows us to empirically evaluate how task complexity and compatibility influence the stability and effectiveness of low-resource multi-task models.

IV. RESULTS

This section addresses the core research questions of our study: (1) how various tasks influence one another during multi-task learning (MTL) under different weight-sharing configurations, and (2) the importance of task selection, as illustrated by incorporating the complex task of two-speaker diarization.

A. Impact of weight-sharing ratio

A general trend observed across all task subsets (Table III and Table IV) is that increasing the proportion of shared parameters correlates with a slight degradation in overall task performance. For instance, in the pair comprising age classification and speech command recognition (SCR), sharing only 7 convolutional layers results in a performance decrease of approximately 2%. This reduction grows to 7% when sharing all 8 convolutional layers, and reaches 10.5% when the attention mechanism is also shared.

However, the extent of accuracy degradation is highly dependent on the specific task set. For instance, the combination of gender classification (GC) and voice activity detection (VAD) exhibited strong resilience to parameter sharing, exhibiting a performance reduction of only 0.4% even under complete sharing of all eight convolutional layers and the attention mechanism. In contrast, the model's performance on speech command recognition (SCR) and age classification (AC) deteriorated significantly under an equivalent sharing configuration, despite the use of task-specific decoders with dedicated attention layers.

This differential sensitivity was further pronounced in three-task configurations. Sets containing both SCR and AC experienced substantial performance degradation under deeper parameter sharing. Conversely, task sets containing only one of these tasks exhibited greater robustness; the combination of VAD, GC, and AC showed a quality reduction of only 3.6% under complete sharing, while the set comprising VAD, GC, and SCR maintained near state-of-the-art performance (a reduction of only 1.3%) even with full sharing of all eight convolutional layers.

Notably, introducing VAD to the AC–SCR pair resulted in a significant quality drop of 7.3% under full sharing including attention. Similarly, adding GC to the same pair led to a 6.3% reduction in performance, even when sharing only the convolutional layers. These results underscore the critical role of task compatibility and the non-linear interactions that arise in multi-task learning, particularly under constrained parameter budgets.

TABLE IV. EXPERIMENTAL RESULTS FOR ALL HIGH-ORDER TASKS SUBSETS

Sharing	Size	VAD	GC[1]	AC[2]	SCR	AR[3]
Partial	57K	92.9	4.8	34.2	92.5	-2.1%
Full CNN	45K	92.8	6.5	37.7	91.1	-7.2%
Complete	43.5K	86.4	14.2	40.5	88.8	-11.5%
Partial	52K	92.2	4.9	33.2		-1.2%
Full CNN	44K	92.6	5.4	33.7		-1.7%
Complete	43.5K	90.6	5.1	35.2		-3.6%
Partial	52K	94.2	4.3		94.5	-0.6%
Full CNN	44K	92.1	4.9		93.7	-1.3%
Complete	43.5K	91.7	9.4		90.4	-5.7%
Partial	52K	92.5		34.6	93.9	-2.6%
Full CNN	44K	92.8		36.5	92.4	-5.5%
Complete	43.5K	90.4		37.7	90.4	-7.3%
Partial	39K		5.4	35.9	92.8	-4.6%
Full CNN	31K		7.1	37	92.5	-6.3%
Complete	29.8K		9.4	39.7	86.7	-10.3%

[1] Gender classification
[2] Age group classification
[3] Relative accuracy reduction

B. Pairwise task synergy

The degree of synergy varies significantly across different task pairs (Table III). The most synergistic pair is voice activity detection (VAD) and gender classification, where the MTL model achieves near state-of-the-art performance—exhibiting a relative accuracy reduction of no more than 0.4%—even in complete sharing mode (8 layers + attention). In contrast, the least synergistic pair is SCR and age classification, where sharing 8 convolutional layers leads to a performance decrease of nearly 7%. Furthermore, VAD emerges as the most universally compatible task in MTL setups, as its performance degrades by less than 2% across all sharing configurations.

A distinct pattern of performance degradation is observed for pairs of global classification tasks. Model accuracy remains near state-of-the-art when sharing 7 or 8 convolutional layers while maintaining task-specific attention mechanisms. However, sharing the attention layer results in significant quality reduction. For instance, for the pair comprising speech command recognition (SCR) and gender classification (GC), performance decreases by merely 1.2% under full CNN sharing, but declines by up to 7.8% when the attention mechanism is shared.

Conversely, the pair of gender classification (GC) and age classification (AC) exhibits strong inherent synergy. This is evidenced by the minimal performance reduction observed even under complete parameter sharing, which includes a single attention layer common to both tasks. This result suggests that these biometric tasks are highly correlated and can be processed simultaneously within a shared representation space with remarkable efficiency, highlighting the importance of task relatedness in multi-task learning performance.

This behavior, however, does not hold for combinations involving voice activity detection (VAD)—a frame-wise classification task—alongside global tasks. We hypothesize that this discrepancy arises from the fundamental differences in output structure between frame-wise and global classification problems. Specifically, the VAD decoder may rely less on attention-based feature refinement, instead leveraging its temporal modeling capacity (e.g., via GRU layers) to achieve high performance. Interestingly, the shared attention layer does not introduce disruptive interference to VAD performance, suggesting that the attention mechanism may remain functionally viable for both task types when appropriately isolated or designed.

These observations highlight the nuanced interplay between architectural sharing and task nature, emphasizing that the optimal sharing strategy must account for both the semantic and structural characteristics of the tasks involved.

C. Higher-order task combinations

We next examine MTL performance on larger task sets (Table IV). For the combination of all four primary tasks, a model sharing 7 convolutional layers achieves performance within 2.1% of the single-task baselines, while reducing the total model size by 53% (utilizing only 57,000 parameters). This result suggests a promising scalable paradigm: adding future tasks may require only one additional convolutional layer with attention and a small task-specific decoder, implying the possibility of solving numerous tasks with drastically reduced resource consumption. However, deeper weight-sharing (e.g., sharing all layers including attention) leads to more substantial quality degradation, with performance reductions of up to 10%.

Analysis of three-task subsets (Table IV) indicates that sharing the attention mechanism consistently degrades performance, likely due to the need for task-specific feature aggregation from the final encoder representation. This architectural choice resulted in performance reductions ranging from 3.6% to 30%. The most effective three-task configuration consists of voice activity detection (VAD), gender classification, and speech command recognition (SCR), which exhibited only a 1.3% performance reduction while achieving a 52% reduction in model size. Interestingly, although age classification appears conceptually similar to gender classification, it demonstrates

TABLE V. PERFORMANCE OF MULTI-TASK MODELS COMBINING SPEAKER DIARIZATION WITH OTHER TASKS. RESULTS SHOW THE DIARIZATION ERROR RATE (DER) FOR THE DIARIZATION TASK AND CORRESPONDING METRICS FOR CO-TRAINED TASKS

Configuration	Sharing	DER	STM[4]	AR[3]
Diar+VAD	Partial	14.7	79.1	-15%
	Full CNN	14.5	82.5	-11.5%
	Complete	14.5	82.5	-11.5%
Diar+GC[1]	Partial	13.5	6.1	-2.4%
	Full CNN	13.7	10.8	-7.3%
	Complete	14.9	18	-15%
Diar+AC[2]	Partial	15	39.5	-12.1%
	Full CNN	15.9	40.4	-19%
	Complete	17.1	42	-37.5%
Diar+SCR	Partial	14	92.7	-4.7%
	Full CNN	16.8	91.5	-25.2%
	Complete	17.1	92.7	-27.3%

[1] Gender classification
[2] Age group classification
[3] Relative accuracy reduction
[4] Second task metric value

notably weaker synergistic properties with other tasks. The most synergistic trio overall proved to be VAD, gender classification, and age classification, which maintained competitive performance with only a 3.6% reduction in overall quality despite employing complete parameter sharing.

D. Effect of adding a complex task

To further analyze the influence of task complexity on multi-task learning performance, we introduced an additional challenging task: two-speaker diarization. This task was implemented using Permutation Invariant Training (PIT) [24] and evaluated on the Libri2Mix dataset [25], with performance measured by the Diarization Error Rate (DER). Our single-task baseline model for diarization achieved a DER of 13.4%.

The incorporation of this complex task starkly illustrates the boundaries of task compatibility within a low-parameter MTL framework. Diarization exhibited negligible synergistic properties with any other task in our set. When trained in a two-task setup alongside any other objective, model performance degraded substantially, with the DER increasing by up to 37.5% (Table V). More critically, introducing diarization to existing synergistic pairs of tasks caused a pronounced negative impact on performance, adversely affecting not only the diarization task itself but also the accuracy of all other co-trained tasks (Table VI).

This result underscores a critical finding: the careful selection of compatible tasks is paramount for successful MTL in resource-constrained environments. It also serves to validate that our identified set of four primary tasks—VAD, SCR, and age and gender classification—constitutes a uniquely synergistic combination, as the introduction of a more complex, incompatible task like diarization severely disrupts the stability and performance of the entire system.

TABLE VI. EXPERIMENTAL RESULTS FOR HIGH-ORDER TASKS SUBSETS INCLUDING DIARIZATION PROBLEM

Sharing	Diar	VAD	GC[1]	AC[2]	SCR	AR[3]
Partial	19.3	91.8	8.2	36.1	93.2	-44%
Full CNN	22.7	87.7	12.2	47.4	88.7	-70%
Complete	23.2	80.3	11.3	42.1	85.5	-72%
Partial	16.1		8		94.2	-20%
Full CNN	14.9		5		92.6	-11.5%
Complete	16.7		12.7		92.6	-24.6%
Partial	14.6			37.4	91.9	-8.7%
Full CNN	14.9			38.3	90.9	-11.3%
Complete	17.2			44.2	93.4	-28.4%

[1] Gender classification
[2] Age group classification
[3] Relative accuracy reduction

V. LIMITATIONS AND FUTURE WORK

While the proposed multi-task learning framework achieves state-of-the-art performance with a significantly reduced parameter footprint compared to an ensemble of single-task models, several limitations warrant discussion and present avenues for future research.

Firstly, this study focuses exclusively on architectural efficiency through weight-sharing and does not incorporate other powerful model compression techniques, such as quantization or pruning. Integrating these methods could potentially yield further reductions in model size and inference latency, enhancing deployability on even more resource-constrained hardware.

Secondly, our investigation is constrained to a specific class of encoder-decoder architectures, namely convolutional neural network (CNN) encoders supplemented with GRU-based decoders for temporal tasks. The generalizability of our findings on task synergy and optimal sharing strategies to other state-of-the-art architectures—such as transformers or more recent convolutional variants like depth-wise separable convolutions—remains an open question.

Thirdly, the training protocol employed a standardized set of hyperparameters across all experimental conditions to ensure a controlled comparison. Consequently, the performance reported herein may not represent the absolute peak achievable for each model configuration. A more extensive, task-specific hyperparameter optimization campaign could potentially lead to further improvements in accuracy and stability.

Finally, the scalability of our framework to a much larger number of tasks (e.g., 10-20) remains unexplored. The current analysis, which includes up to four tasks, suggests that task compatibility is crucial. It is plausible that adding a multitude of tasks, particularly those that are complex or mutually antagonistic, could lead to increased negative interference and performance degradation, challenging the robustness of the proposed sharing strategies. Future work will involve stress-testing the architecture's capacity and developing more dynamic weight-sharing mechanisms to accommodate larger and more diverse task sets.

VI. CONCLUSIONS

This paper presented a comprehensive study on low-complexity multi-task learning for audio analysis, addressing

two fundamental challenges: optimal weight-sharing strategies and task compatibility assessment. Our key contributions are threefold. First, we demonstrated that partial weight-sharing (7 convolutional layers) enables effective integration of four core audio tasks—voice activity detection, speech command recognition, and age/gender classification—while reducing model size by 53% (57K parameters) compared to single-task ensembles. Second, our systematic analysis revealed distinct task synergy patterns, showing that voice activity detection exhibits strong compatibility with speaker biometrics, while speech command recognition and age classification require more careful integration. Third, we established that introducing complex tasks like speaker diarization can cause substantial performance degradation, highlighting the critical importance of strategic task selection. These findings provide practical guidelines for designing efficient multi-task audio systems suitable for resource-constrained edge devices.

ACKNOWLEDGMENT

I extend my deepest gratitude to my beloved wife, Daria Tarasova, for her invaluable contributions to the manuscript review process and insightful discussions.

REFERENCES

[1] R. Hebbar, K. Somandepalli, and S. Narayanan, "Robust speech activity detection in movie audio: Data resources and experimental evaluation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4105–4109.

[2] M. Sharma, S. Joshi, T. Chatterjee, and R. Hamid, "A comprehensive empirical review of modern voice activity detection approaches for movies and tv shows," *Neurocomputing*, vol. 494, pp. 116–131, 2022.

[3] D. C. De Andrade, S. Leo, M. L. D. S. Viana, and C. Bernkopf, "A neural attention model for speech command recognition," *arXiv preprint arXiv:1808.08929*, 2018.

[4] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3535–3552, 2022.

[5] Khaled Koutini and Jan Schlüter and Hamid Eghbal-zadeh and Gerhard Widmer, "Efficient Training of Audio Transformers with Patchout," in *Interspeech 2022*, 2022, pp. 2753–2757.

[6] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 5178–5193.

[7] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2024.

[8] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Interspeech 2021*, 2021, pp. 3665–3669.

[12] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[9] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6167–6171.

[10] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[11] M. Ayache, H. Kanaan, K. Kassir, and Y. Kassir, "Speech command recognition using deep learning," in *2021 sixth international conference on advances in biomedical engineering (ICABME)*. IEEE, 2021, pp. 24–29.

[13] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE transactions on knowledge and data engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[14] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[15] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 2014, pp. 94–108.

[16] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3150–3158.

[17] X. Zhao, H. Li, X. Shen, X. Liang, and Y. Wu, "A modulation module for multi-task learning with applications in image retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–416.

[18] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1871–1880.

[19] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1930–1939.

[20] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3994–4003.

[21] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, "Latent multi-task architecture learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 4822–4829.

[22] D. A. Krause and A. Mesaros, "Binaural signal representations for joint sound event detection and acoustic scene classification," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 399–403.

[23] T. Khandelwal and R. K. Das, "A multi-task learning framework for sound event detection using high-level acoustic characteristics of sounds," in *Interspeech 2023*, 2023, pp. 1214–1218.

[24] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[25] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation."