Evaluating and Interpreting Pooling Techniques in Spectrogram-Based Audio Analysis Using Diverse Metrics

Supun Bandara, Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa, Katubedda, 10400, Sri Lanka

***Abstract*—Audio analysis is a rapidly advancing field that spans various domains, including speech, music, and environmental sound data. Using spectrograms with Convolutional Neural Networks (CNNs) enables the visualization and extraction of critical audio features by combining time-frequency representations with deep learning. Pooling plays a crucial role in this process, as it reduces dimensionality while retaining essential information. However, existing evaluations of pooling methods primarily emphasize downstream task performance, such as classification accuracy, often overlooking their effectiveness in preserving critical signal features. To address this gap, we use 17 distinct metrics, categorized into four domains, to comprehensively assess various pooling operations. Furthermore, we explore the underexamined relationship between specific pooling techniques and their impact on feature retention across diverse audio applications. Our analysis encompasses spectrograms from three audio domains (speech, music, and environmental sound), identifying their key characteristics, and grouping them accordingly. Using this setup, we evaluate the performance of 12 pooling methods across these applications. By investigating the features critical to each task and evaluating how well different pooling techniques preserve them, we give insights into their suitability for specific applications. This work aims to guide researchers in selecting the most appropriate pooling strategies for their applications, enabling more granular evaluations, improving explainability, and thereby advancing the precision and efficiency of audio analysis pipelines.**

***Keywords*—Audio data analysis; pooling; deep learning; dimensionality reduction; spectrograms**

I. Introduction

Audio analysis has become an essential field of study with wide-ranging applications, including audio classification [1], [2], [3], speaker recognition [4], [5], [6], and sound event detection [7], [8], [9], [10], [11]. The ability to accurately process and analyze audio data is crucial for developing systems that can understand and interact with human environments. Over the past few decades, significant advances in machine learning (ML) and deep learning have revolutionized audio analysis, enabling more sophisticated and accurate methodologies.

Feature extraction is a critical step in the analysis of audio signals, where audio signals are transformed into representations that can effectively capture meaningful patterns and features. Over the years, a wide range of techniques have been developed to extract meaningful information from audio data [12]. Time-domain approaches analyze raw waveforms to derive features such as zero-crossing rate and signal energy [12], while frequency-domain methods, including the Fourier Transform [13] and Mel-Frequency Cepstral Coefficients (MFCCs) [14], focus on spectral characteristics. Time-frequency representations, such as spectrograms [15] and chromagrams [16], combine the strengths of both domains, capturing temporal dynamics and spectral content simultaneously. Recent advances in ML have further introduced data-driven methods, such as Convolutional Neural Networks (CNN) [17], Recurrent Neural Networks [18], and transformers [19], which learn task-specific features directly from raw audio or time-frequency representations.

By combining traditional signal processing approaches with advanced deep learning methods, these techniques can enable robust and scalable solutions, driving progress in diverse audio applications and opening the door to innovative research and development. One of those approaches is using spectrograms with CNNs. A significant advantage of using this lies in their ability to harness the strengths of both time-frequency representations and deep learning models [20], [21]. Spectrograms provide a rich and visually interpretable representation that captures both temporal dynamics and spectral patterns of audio signals. CNNs excel at learning hierarchical and spatial features, enabling the identification of complex harmonic structures, frequency modulations, and localized spectral features in the spectrogram. Despite their effectiveness, the integration of spectrograms with CNNs presents challenges due to its high dimensionality, which increases computational demands and memory usage [22]. To address this issue, pooling methods are employed to reduce the dimensionality of spectrograms while preserving essential information. Pooling techniques such as max pooling [1] and average pooling [2] summarize spectral content, thereby improving computational efficiency and enabling the training of deep learning models on large audio datasets.

Traditionally, the effectiveness of pooling methods in audio analysis has been evaluated indirectly using the accuracy of the downstream tasks such as classification [23]. However, this approach may not fully capture how well pooling methods preserve critical features within spectrograms. In this study, we apply 17 evaluation metrics categorized into four distinct groups for evaluating pooling methods used in audio analysis. These metrics can be used to directly assess the ability of pooling methods to preserve critical features and patterns in spectrograms, independent of downstream task results. By focusing on the intrinsic performance of pooling techniques, this approach enables a more granular analysis of their impact on the representation of audio features. Our investigation spans

12 pooling methods, facilitating a standardized comparison of these techniques. This analysis gives insights to identify the pooling methods best suited to specific applications by investigating which audio features should be preserved for specific applications and how well different pooling methods preserve these features. This work provides insights into optimizing the pooling process, helping researchers better understand the features relevant to their applications, and improve accuracy by identifying better pooling methods for their applications. Therefore, it facilitates better model design, improves the explainability of deep audio systems, and supports more informed decisions in the deployment of audio analysis pipelines for real-world applications.

The rest of the study is organized as follows: Section II provides the theoretical background, including a detailed review of pooling methods and evaluation metrics. Section III describes the system architecture. Section IV presents the experimental setup and includes an in-depth spectrogram analysis. Section V outlines the experimental findings across multiple audio domains, highlighting the quantitative performance of various pooling methods, and Section VI interprets the experimental findings, linking observed patterns to the functional demands of each audio task. Section VII concludes the study, highlighting key findings and directions for future work.

II. Theoretical Background

This section outlines the theoretical foundations of spectrogram-based audio analysis, focusing on the characteristics of speech, music, and environmental sounds. It introduces key pooling methods used to reduce spectrogram dimensionality while preserving important features. To assess their effectiveness, we also present evaluation metrics that measure information retention, structural preservation, local detail, and compression, forming a basis for analyzing pooling performance across audio tasks.

A. Local Details and Global Structure of a Spectrogram

Audio signals can generally be categorized into three primary types: speech, music, and environmental sounds [12].

1) Speech: Speech is generated by humans through the combined activity of organs such as the lungs, vocal cords, mouth, nose, and brain. The vocal cords and vocal tract play a crucial role in shaping speech sounds [24]. The frequency range of human speech typically begins at 100 Hz and can extend up to 17 kHz.

2) Music: Musical sounds are created by instruments or the human voice to produce harmony and express emotions. Music can be analyzed based on characteristics such as genre, mood, and tonal quality [25]. Traditional genres include rock, jazz, classical, and pop. The frequency range of music usually falls between 40 Hz and 19.5 kHz.

3) Environmental sounds: The sounds we encounter daily, such as those from vehicles, running water, doorbells, phones, machinery, and animals, are classified as environmental sounds [26]. These sounds often cover the full range of audible frequencies.

Fig. 1 displays the time-domain waveforms of three example sound types: human speech, a guitar note, and a car horn. Speech and music signals exhibit periodicity, whereas environmental sounds generally lack any clear periodic pattern. As shown in Fig. 1, speech signals are typically smooth and continuous, while guitar notes are brief and intermittent. In comparison, sounds such as a fire truck siren appear as noisy signals with high amplitude. Differences between these sounds are also apparent in the time-frequency domain, as illustrated in Fig. 2, which highlights the distinct frequency spectra of each sound type.

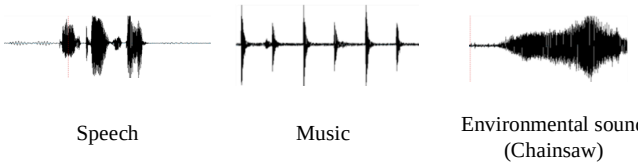

Fig. 1. Time-domain waveforms for speech, music, and environmental sound data.

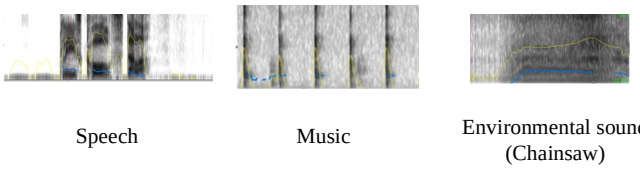

Fig. 2. Time-frequency domain waveforms for speech, music and environmental sound data.

By creating spectrograms for these three audio categories, we can analyze the features that need to be preserved during processing. This analysis is crucial when applying pooling or other transformations to spectrograms, as it helps ensure that the relevant features are retained in the pooled outputs. By analyzing spectrograms, we can determine what features to preserve for different applications. Based on the analysis, these applications can be categorized into two distinct groups: global structure preservation and localized detail retention. This categorization is driven by the need to either retain overall patterns in the spectrogram or emphasize specific localized features.

B. Review of Pooling Methods

Pooling techniques are important to audio analysis, offering a way to condense critical information over time or frequency dimensions. By summarizing features, pooling enhances the robustness of models against variations in speech signals, such as differences in speaking rates, accents, and background noise. This section explores a range of pooling methods, detailing their roles in reducing dimensionality while retaining essential features.

1) Average pooling: Average pooling [2], [27] computes the mean value of elements in a pooling region, and is expressed as in Eq. (1).

$$f_{\text{avg}}(x) = \frac{1}{N} \sum_{i=1}^{N} |x_i| \tag{1}$$

This method smooths feature maps by averaging values, effectively reducing noise. However, since it assigns equal weight to all elements, background regions can dominate, potentially diminishing the model's ability to distinguish between critical and irrelevant features.

2) Max pooling: Max pooling [1], [28], [29] selects the maximum value from a pooling region. Eq. (2) shows how max pooling can be used in a feature map.

$$f_{\max}(x) = \max\{x_i\}_{i=1}^{N} \tag{2}$$

This approach highlights the strongest activations, reducing the influence of background noise and emphasizing prominent features. Despite these advantages, max pooling can overlook subtle but important information and may amplify noisy elements with high values.

3) Min pooling: Min pooling [30] extracts the smallest value in a pooling region, as shown in Eq. (3).

$$f_{\min}(x) = \min\{x_i\}_{i=1}^{N} \tag{3}$$

By focusing on the weakest feature in each region, min pooling suppresses noise and outliers, making it useful in tasks such as anomaly detection. However, it shares limitations with max pooling, such as losing mid-range information and subtle details.

4) Mixed pooling: Mixed pooling [31] combines the strengths of max and average pooling by using a weighted combination as defined in Eq. (4).

$$f_{\text{mixed}}(x) = \alpha \cdot f_{\max}(x) + (1-\alpha) \cdot f_{\text{avg}}(x) \tag{4}$$

The parameter α determines the balance between max and average pooling. For $\alpha = 1$, it behaves as max pooling, and for $\alpha = 0$, it becomes average pooling. Mixed pooling can adapt to different tasks and datasets by capturing both salient features and broader contextual information.

5) Linear softmax pooling: Linear softmax pooling (LSP) [32] assigns weights to features based on their squared values, ensuring prominent features receive higher emphasis, as defined in Eq. (5).

$$f_{LSP} = \frac{\sum_i x_i^2}{\sum_i x_i} \tag{5}$$

where, f_{LSP} represents the output of the softmax function and x_i represents the individual elements of the input vector.

This method is particularly effective in scenarios, where identifying dominant peaks in data distributions is critical, as it balances emphasis on key features while considering all inputs.

6) Exponential softmax pooling: Exponential softmax pooling (ESP) [33] uses an exponential transformation to accentuate differences between input features in accordance with Eq. (6).

$$f_{ESP} = \frac{\sum_i x_i \exp(x_i)}{\sum_i \exp(x_i)} \tag{6}$$

where, $\exp(x_i)$ denotes the exponential function applied to the individual elements x_i in the input vector. By emphasizing higher values, this approach excels in tasks requiring sharp distinctions between features, such as identifying critical events in a data sequence.

7) Learned-norm pooling: As described in Eq. (7), learned-norm pooling (LNP) [34] unifies max and average pooling principles using the p-norm.

$$f_{LNP}^{p}(x) = \left(\frac{1}{n}\sum_{i=1}^{n}|x_i|^p\right)^{1/p} \tag{7}$$

The parameter p, ranging from 1 to infinity, determines the behavior of the pooling. When $p = 1$, it corresponds to mean pooling, while $p \to \infty$ approximates max pooling. LNP provides flexibility in capturing nuanced feature distributions.

8) Log-Sum-Exp pooling: Log-Sum-Exp Pooling(LSEP) [35] offers a smooth interpolation between mean and max pooling. This behavior is captured by Eq. (8).

$$f_{LSEP}^{r}(x) = \frac{1}{r}\log\left(\frac{1}{n}\sum_{i=1}^{n} exp(r \cdot x_i)\right) \tag{8}$$

The parameter r controls the pooling behavior, with $r \to 0$ resembling mean pooling and $r \to \infty$ approximating max pooling.

9) Auto-pooling: Auto-pooling [33] adapts dynamically to data, combining max and average pooling principles as formulated in Eq. (9).

$$f_{auto}^{\alpha} = \frac{\sum_j \mathbf{x}_j \exp(\alpha \mathbf{x}_j)}{\sum_j \exp(\alpha \mathbf{x}_j)} \tag{9}$$

This adaptability is particularly useful in sound event detection, where event duration and intensity vary significantly. Auto-pooling excels in environments with diverse and unpredictable audio patterns.

10) Power pooling: Power pooling [36], [37] introduces a trainable parameter n to adjust frame-level predictions, as defined in Eq. (10).

$$f_{power}^{c} = \frac{\sum_i x_i^f \times (x_i^f)^n}{\sum_i (x_i^f)^n} \tag{10}$$

This method transitions between max and average pooling, making it suitable for applications involving weakly labeled datasets or varying temporal scales.

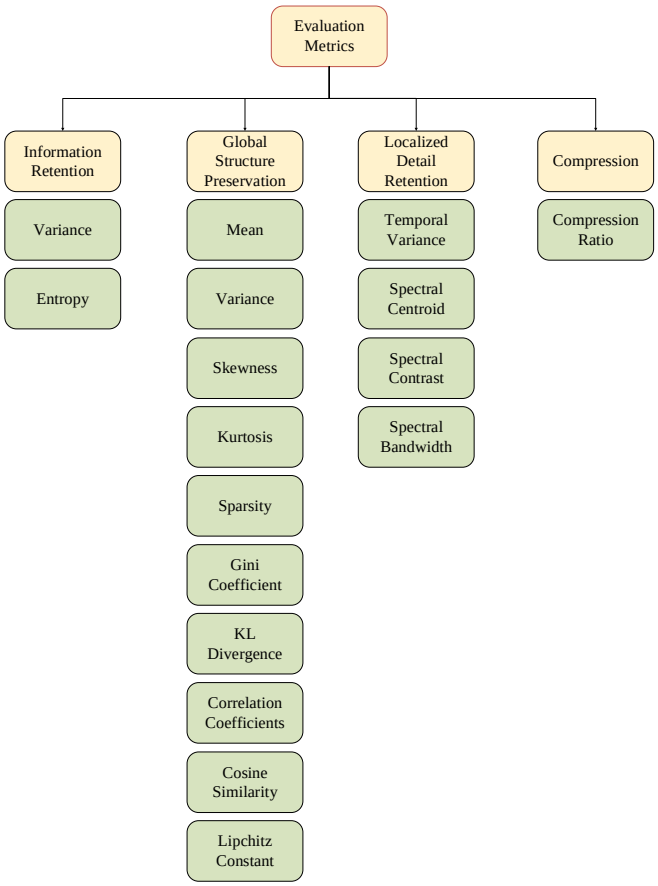

Fig. 3. Types of evaluation metrics categorized based on different usages.

11) Entropy pooling: Entropy pooling [38] selects high-entropy features, ensuring the most informative elements are retained. Eq. (11) demonstrates how entropy pooling works.

$$f_{\text{entr}}(X_r) = X_r[g(P_r)] \tag{11}$$

where, f_{entr} is the entropy-pooled output. By ensuring that the pooled output retains the most informative and least redundant features, entropy pooling reduces redundancy and maintains a more uniform feature distribution, thereby making the model more resilient to noise.

12) Attention pooling: Attention pooling [39], [40], [4], [41], [5], [42] dynamically assigns weights to input features, focusing on the most relevant parts of the audio signal. It uses a learnable weight vector $\mathbf{w}_i$ to emphasize frames critical for the task, creating a weighted average output.

C. Metrics for Evaluating Pooled Spectrograms

We have identified a set of metrics that can be used to evaluate the performance of pooling methods applied to spectrograms. Based on their specific focus, Fig. 3 shows how these metrics can be grouped into four categories: information retention, preservation of global structure, retention of localized details, and dimensionality reduction. These metrics ensure that the pooled spectrograms remain representative of the original data while achieving dimensionality reduction.

1) Information retention: Information retention metrics evaluate whether the pooled spectrogram preserves sufficient variability, complexity, or detail from the original spectrogram. These metrics ensure that pooling does not discard too much essential information.

a) Variance: Variance [43] quantifies the spread of intensity values in the spectrogram. As shown in Eq. (12), it measures how much the intensity values deviate from their mean, representing the variability in the data. In the context of pooled spectrograms, preserving variance ensures that dynamic range and signal variability are maintained. A significant drop in variance after pooling could indicate a loss of essential details, such as transitions or contrasts in intensity.

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{12}$$

where, x_i is intensity value at pixel i in the spectrogram, μ is mean of all intensity values ($\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$), and N is total number of pixels or intensity values.

b) Entropy: Entropy [44] assesses the complexity or randomness of the spectrogram by evaluating the distribution of intensity values, as given in Eq. (13). Higher entropy indicates a richer or more diverse representation of information. If the entropy of the pooled spectrogram is close to the original, it implies that essential patterns and nuances, such as background noise or intricate harmonic content, are preserved. A decrease in entropy suggests that the pooling operation may have oversimplified the spectrogram, discarding finer details.

$$\text{Entropy} = -\sum_{i=1}^{N} p(x_i)\log p(x_i) \tag{13}$$

where, $p(x_i)$ is the probability of intensity value x_i in the spectrogram, and N is the total number of intensity levels.

2) Global structure preservation: Global structure preservation metrics evaluate whether the structural and temporal patterns of a spectrogram are retained after pooling, which is crucial for applications that rely on macro-level spectrogram features. Pooling reduces the size of a spectrogram, but maintaining these patterns ensures that the essential characteristics remain intact. Several types of pattern preservation metrics can be identified, focusing on aspects such as statistical properties, near-zero values, energy distribution, and similarity measures.

a) Statistical properties: Preserving statistical properties ensures that the overall patterns and distributions in the spectrogram remain consistent after pooling [45]. Key statistical metrics include,

- Mean [45], which represents the average intensity of the spectrogram. Preserving the mean ensures that the global energy level is consistent, reflecting the overall sound intensity. Mean is calculated as in Eq. (14)

$$\text{Mean} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{14}$$

where, x_i is the intensity value at pixel i, and N is the total number of pixels.

- Variance [45], which captures the spread of intensity values, preserving the variability and dynamic range of the spectrogram.
- Skewness [45], which measures asymmetry in the intensity distribution, as presented in Eq. (15). Maintaining skewness ensures that pooling does not disproportionately favor high-energy or low-energy regions.

$$\text{Skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3} \tag{15}$$

where, x_i is the intensity value at pixel i, μ is the mean of intensity values, and σ is the standard deviation ($\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$).

- Kurtosis [45], which reflects the sharpness or flatness of the intensity distributions, in accordance with Eq. (16). Preserving kurtosis retains critical features such as peaks (high energy areas) or flat regions (low activity areas).

$$\text{Kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \tag{16}$$

where, x_i is the intensity value at pixel i, μ is the mean of intensity values, and σ is the standard deviation. Together, these metrics ensure that the original statistical characteristics of the spectrogram are reflected in the pooled version.

b) Low activity regions: Sparsity [46] refers to the proportion of near-zero or inactive regions in a spectrogram, and retaining it is essential in applications, where silent or low-activity regions convey meaningful information. For instance, in speech recognition, silent gaps or pauses are crucial for segmenting phonemes or words, while in speaker recognition, low-energy regions may contain speaker-specific traits. Pooling methods must preserve these sparse areas to ensure critical contextual information is not lost. Sparsity is calculated as shown in Eq. (17), and sparsity ratio, which is given in Eq. (18), is calculated in our experiments.

$$\text{Sparsity} = \frac{\text{Number of near-zero values}}{\text{Total number of values}} \tag{17}$$

$$\text{Sparsity ratio} = \frac{\text{Sparsity of pooled spectrogram}}{\text{Sparsity of original spectrogram}} \tag{18}$$

c) Energy distribution: The Gini coefficient [47] measures the inequality in intensity distribution across a spectrogram, where a lower value indicates a more uniform distribution and a higher value reflects concentrated energy, such as dominant frequency bands. By comparing the Gini coefficients before and after pooling, it is possible to evaluate whether the pooling process has preserved the energy balance across the spectrogram.

d) Similarity measures: Similarity measures directly compare the pooled spectrogram with the original to ensure that key structural and temporal features are preserved. Some of the similarly metrics which might be useful for evaluating the global structure preservation are as follows:

- Kullback-Leibler (KL) Divergence [48] quantifies the difference between intensity distributions, according to Eq. (19). A lower divergence indicates that the pooled spectrogram closely resembles the original in terms of energy patterns.

$$D_{\text{KL Divergence}}(P|Q) = \sum_{i=1}^{N} p(x_i) \log \frac{p(x_i)}{q(x_i)} \tag{19}$$

where, $p(x_i)$ is the probability distribution before pooling, and $q(x_i)$ is the probability distribution after pooling.

- Correlation Coefficients [49], [50] evaluates the linear relationship between intensity values in the original and pooled spectrograms, ensuring that proportional changes are consistent.
- Cosine Similarity [51] assesses alignment between feature vectors of the original and pooled spectrograms, indicating the degree to which proportional feature retention is achieved, as given in Eq. (20).

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} \tag{20}$$

where, x_i is the intensity value in the original spectrogram, and y_i is the intensity value in the pooled spectrogram.

- Lipschitz Continuity [52] ensures that small changes in the original spectrogram correspond to small changes in the pooled spectrogram, preventing distortions that could compromise structural integrity [see Eq. (21)].

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\| \tag{21}$$

where, $f(x)$ is the mapping after pooling, and L is the Lipschitz constant.

3) Localized detail retention: Feature-specific metrics assess the retention of localized or detailed aspects of the spectrogram. These are crucial for fine-grained tasks such as instrument identification or event detection.

a) Temporal features: Temporal features [53] focus on intensity fluctuations, transitions, and coherence over time. As shown in Eq. (22), temporal variance is a metric that measures how well time-based variations, such as rhythm or speech patterns, are preserved. Temporal feature retention is vital for applications, where timing and transitions play a critical role, such as in rhythm analysis or speech recognition.

$$\text{Temporal Variance} = \frac{1}{T}\sum_{t=1}^{T}(I_t - \mu_t)^2 \tag{22}$$

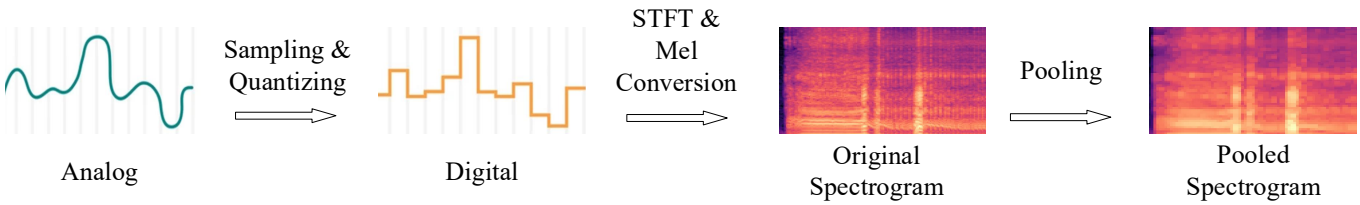

Fig. 4. Illustration of the feature extraction and pooling process of audio analysis applications. Evaluation metrics are used to compare the original and pooled spectrograms.

where, I_t is the intensity at time t, μ_t is the mean intensity over time, and T is the total number of time frames.

b) Spectral features: Spectral metrics evaluate the preservation of frequency-based characteristics. They are calculated as in respective reference papers.

- Spectral Centroid [54]: Represents the "center of mass" of the frequency spectrum, indicating the dominant frequency range. Preserving the spectral centroid ensures that the tonal characteristics remain consistent.
- Spectral Contrast [55]: Captures the differences between peaks and valleys in the frequency spectrum. Retaining spectral contrast ensures the preservation of harmonic content and overall tonal richness.
- Spectral Bandwidth [54]: Measures the spread of frequencies in the spectrum. A high spectral bandwidth indicates a broad frequency range, while a low bandwidth suggests focused frequency content. Preserving bandwidth ensures that the spectral characteristics of the pooled spectrogram match the original.

4) Compression: Compression-focused metrics [56] measure the effectiveness of pooling methods in reducing spectrogram dimensions while preserving essential information. A key metric is the compression ratio, which quantifies the extent of dimensionality reduction and reflects the balance between minimizing size and retaining critical features. The compression ratio is calculated as given in Eq. (23). An effective pooling method achieves this balance by eliminating redundancy without compromising the spectrogram's informational content.

$$\text{Compression Ratio} = \frac{\text{Original size of spectrogram}}{\text{Pooled size of spectrogram}} \tag{23}$$

III. System Architecture

In Fig. 4, the steps used for generating a spectrogram from an audio file are outlined. The process starts with sampling the analog audio signal at a fixed rate of 16 kHz and quantizes the samples into discrete values. The digital signal is divided into overlapping frames, and the Hann windowing function is applied to reduce spectral leakage. Each frame undergoes a Short-Time Fourier Transform (STFT) to convert the time-domain signal into a frequency-domain representation, resulting in a magnitude spectrum. This spectrum is then mapped to the Mel scale to create a mel spectrogram, aligning the frequency bins perceptually. The resulting power spectrogram is converted to a logarithmic scale by normalizing the intensity relative to the maximum value, producing a log-scaled Mel spectrogram suitable for visualization and further analysis. Once the spectrogram is generated from the audio signal, pooling can be applied to the spectrogram itself.

IV. Experiments

A. Experimental Setup

We selected a range of audio classification tasks for our experiments, encompassing all three major audio types: speech, music, and environmental sounds.

1) Speech: Speech Emotion Recognition task is used for the speech data domain. IEMOCAP dataset [57], which contains approximately 10,000 speech utterances labeled with emotions, and focuses on four emotion classes (angry, happy, sad, neutral), is used.

2) Music: For Music Genre Classification, GTZAN dataset [58] is used, which consists of 1000 music excerpts (30 seconds each) spanning 10 musical genres. IMRAS dataset [59] is used for Music Instrument Recognition task.

3) Environmental sounds: Environmental Sound Classification task is used for this domain with ESC-50 dataset [60], which consists of 2000 five-second audio clips in 50 classes. It contains five major categories of animals, natural soundscapes and water sounds, human non-speech sounds, domestic sounds, and exterior sounds, each containing 10 equally balanced classes of sound events.

These datasets were chosen for their standard usage in the respective domains, diversity of audio samples, and relevance to the identification of global versus localized features in spectrogram-based audio analysis. Spectrograms were generated as described in Section III, and analyzed to identify both local and global features. Subsequently, pooling methods described in Section II-B were applied, and the evaluation metrics from Section II-C were used to assess how well each pooling method preserves local and global features. These metrics also help determine which are most effective for measuring feature preservation in spectrogram-based models. Mixed pooling is non-trainable and includes hyperparameters, with its mixing proportion set to a fixed value of 0.5.

B. Analyzing Spectrograms

Spectrograms were generated for each audio type to facilitate visual identification of their characteristic features.

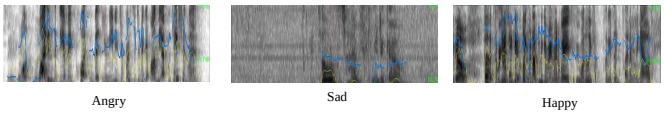

Fig. 5. Voice patterns of the same person with different emotions.

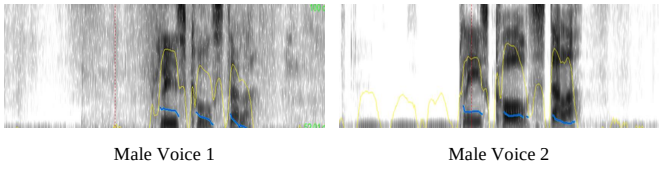

Fig. 6. Voice patterns of different people of same gender talking the same sentence shown in a spectrogram.

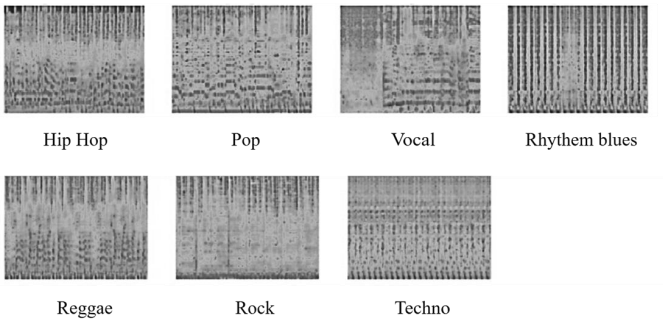

Fig. 8. Spectrogram patterns of different music genres that are used for music genre classification.

1) Speech applications: In speech applications, preserving localized features in the spectrogram is crucial for tasks such as speech emotion recognition, where subtle variations in pitch, intensity, and temporal dynamics carry significant meaning. As shown in Fig. 5, we have identified how these features appear as localized changes in the spectrogram, such as brief intensity surges or shifts in frequency, which are essential for identifying emotional states. For instance, emotions such as anger or happiness are often expressed through louder speech with elevated pitch, while sadness is characterized by softer, flatter intonation patterns. The retention of these intricate features enables a deeper understanding of vocal expressions, making them vital for emotion recognition systems that rely on the nuanced analysis of audio signals.

However, in most speech applications, it is crucial to preserve the overall patterns of the spectrogram. These patterns are influenced by various factors, such as the speaker's unique voice timbre, pitch variations, and speaking style, including how pauses are placed between words. Such variations are particularly significant in speaker recognition tasks, where identifying an individual speaker depends on recognizing unique timbre and consistent frequency patterns over time. For instance, as shown in Fig. 6, two male speakers saying the same sentence exhibit noticeable differences in pauses, frequency patterns, pitch, and intensity variations. Similarly, Fig. 7 highlights the distinctions between male and female speakers for the same sentence. Therefore, it is obvious that when applying pooling, overall patterns of the spectrograms should be preserved in most of the speech applications.

2) Music applications: In music-related tasks, preserving spectrogram patterns is crucial for capturing the intricate details of musical compositions and maintaining audio fea-

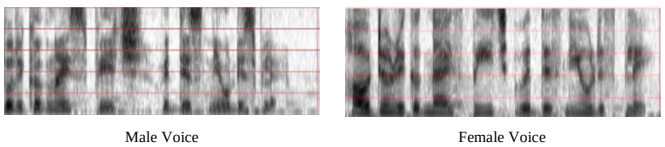

Fig. 7. Voice patterns of male and female voice with the same sentence shown in a spectrogram.

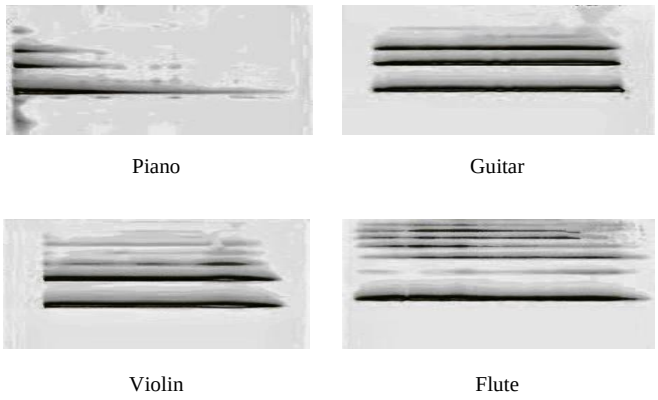

Fig. 9. Spectrogram patterns with different music instruments that is used for music instruments identification.

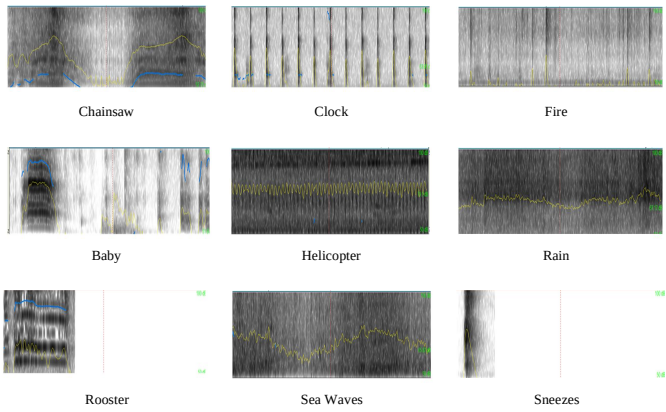

Fig. 10. Spectrogram patterns with different environmental sounds.

tures critical for accurate analysis and classification. We have illustrated in Fig. 8 how different music genres exhibit distinct continuous patterns, which must be preserved to accurately identify those genres. Music genre classification relies on distinguishing broad musical features such as rhythm patterns, harmonic structures, and timbre, each of which reflects unique spectrogram characteristics. Classical music often presents smooth, sustained harmonic patterns, while electronic dance music (EDM) features repetitive beats and sharp intensity changes. These preserved spectrogram features reveal the temporal and spectral patterns necessary for precise classification.

However, as shown in Fig. 9, some music applications such as musical instrument identification depends on preserving timbral qualities, harmonic structures, and frequency ranges unique to each instrument. A piano, for instance, displays evenly spaced horizontal harmonic lines, while a guitar shows richer textures due to its plucked strings. These features, which are often small but distinct, provide critical cues for accurate identification. Additionally, capturing the unique spectral patterns of each instrument, such as the smooth overtones of a violin or the sharp attacks of a drum, is essential for ensuring that their individual characteristics are retained during audio analysis, enabling precise classification in diverse musical contexts.

3) Environmental sound applications: In environmental sound applications, detecting and classifying sounds relies heavily on preserving their distinct temporal and spectral characteristics, which are often brief but highly specific. Environmental sound classification involves recognizing natural sounds such as rain, wind, or traffic noise by retaining their defining spectral features and temporal dynamics such as rain's high-frequency noise bursts or wind's diffuse, broad frequency patterns ensuring these localized elements serve as a foundation for accurate classification. We have highlighted these distinct patterns in Fig. 10 to demonstrate the importance of spectrogram feature preservation in capturing the nuances of various environmental sounds, enabling precise classification across diverse audio tasks.

V. Results

A. General Trends Across Pooling Methods

Across all three domains, several consistent patterns are observed in the behavior of pooling methods. Max pooling consistently preserves the highest variance, mean, skewness, and temporal variance, indicating its strong capacity to retain localized, high-energy features such as transient peaks and short-duration events. While this makes max pooling particularly effective for tasks where such localized features are essential, it often results in the loss of broader structural coherence, as seen in its comparatively lower cosine similarity and correlation coefficients. In contrast, average pooling demonstrates reliable global structure preservation across all datasets. It maintains constant mean retention, cosine similarity, and correlation coefficients, suggesting that it effectively captures the general shape of spectrograms while smoothing out noise and minor variations. This makes average pooling especially suitable for tasks where the continuity of harmonic or rhythmic patterns is more critical than localized precision.

Entropy pooling and ESP consistently exhibit the highest entropy and kurtosis values, reflecting their ability to capture a wide diversity of information. However, these methods simultaneously suffer from poor performance in structure-preserving metrics such as correlation, Lipschitz continuity, and temporal variance, indicating that they degrade interpretability and coherence. This trade-off suggests that while entropy-based methods may retain information richness, they do so at the expense of structural and temporal fidelity.

Auto pooling, mixed pooling, and LSEP emerge as the most balanced approaches. These methods perform well across a broad range of metrics, combining moderate-to-high variance preservation with strong similarity and temporal coherence. Their capacity to retain both global structures and localized features renders them especially well-suited to tasks that require a blend of both information types. Notably, similarity metrics such as cosine similarity and KL divergence show relatively little variation across pooling methods, suggesting that while useful for general alignment assessment, these metrics are less effective for differentiating nuanced pooling behavior compared to statistical or spectro-temporal measures.

B. Task-Specific Observations

1) Speech emotion classification: Speech emotion classification is primarily a localized feature-dependent task. Emotional states are conveyed through subtle and transient modulations in pitch, intensity, and timing, which necessitate the preservation of fine-grained temporal and spectral features. As shown in Table I, max pooling is notably effective in this domain, producing the highest variance (1.1031) and skewness, and excelling in capturing rapid vocal changes associated with expressive emotions such as anger or surprise. However, its emphasis on energetic extremes may occasionally overamplify variations, leading to potential imbalances in emotion recognition.

2) Music genre classification: In music genre classification, the primary analytical focus lies in preserving global spectro-temporal structures such as rhythm, harmony, and tonal balance across longer time scales. As shown in Table II, average pooling consistently performs well in this domain, achieving high mean preservation (0.9988), cosine similarity (0.9922), and correlation (0.9854), making it effective for capturing the overarching structure of musical compositions, particularly in genres characterized by smooth and continuous transitions (e.g., classical, ambient).

3) Environmental sound classification: While environmental sound classification relies on both global and local feature preservation, it primarily depends on local features. Audio scenes in this domain frequently combine ambient background noise with short, high-intensity acoustic events. Table III shows that the effectiveness of max pooling in ESC is evident from its superior variance (1.0477) and temporal variance (1.00), confirming its utility in capturing brief, salient events such as alarms or door slams. However, average pooling also plays a vital role, achieving the highest cosine similarity (0.9876) and correlation coefficient (0.9812), indicating that it effectively preserves background structure and overall scene continuity.

TABLE I. EVALUATING THE PERFORMANCE OF POOLING METHODS ON THE SPEECH EMOTION CLASSIFICATION TASK USING EVALUATION METRICS TO IDENTIFY WHICH FEATURES EACH POOLING METHOD PRESERVES. THE VALUES OF THE POOLED SPECTROGRAM TO THE ORIGINAL SPECTROGRAM RATIOS OR SIMILARITY SCORES ARE GIVEN IN THE TABLE

Pooling Method	Varia. Preser.	Mean Preser.	Skew.	Kurt.	Entropy	Spars. Level	Gini Coeff.	KL Diver.	Cosi. Simil.	Corr. Coef.	Lips. Cont.	Temp. Var.	Spec. Cent.	Spec. Cont.	Spec. Band.	Comp. Ratio
average	0.7120	0.9780	0.0911	2.9371	1.3185	0.6410	-0.0704	28.1342	0.9811	0.9732	0.61	0.88	0.86	0.82	0.83	4.1
max	1.1031	1.1224	0.6550	4.6204	1.4869	0.0048	-0.1003	28.0427	0.9620	0.9540	0.93	1.00	1.12	1.26	1.15	4.3
min	0.6028	0.8393	-0.4835	2.1044	1.2910	0.8724	-0.0582	28.2438	0.9367	0.9273	0.36	0.63	0.74	0.61	0.70	4.0
mixed	0.8611	0.9143	0.2489	3.2649	1.3341	0.0190	-0.0922	27.9881	0.9505	0.9405	0.72	0.90	0.96	0.89	0.93	4.2
LSP	0.7803	1.0122	0.1348	3.0715	1.3742	0.0029	-0.0816	28.1605	0.9699	0.9604	0.59	0.82	0.91	0.84	0.87	4.1
ESP	0.0094	0.0015	-1.9341	10.2145	4.5987	0.4042	0.8840	28.1710	0.0084	0.0103	0.03	0.18	0.31	0.24	0.28	3.2
LNP	0.6812	-0.8841	-1.6221	5.3648	1.3372	0.0340	0.1012	28.0347	0.9350	0.9231	0.28	0.55	0.69	0.62	0.67	4.0
LSEP	1.0225	0.9011	0.4082	3.5653	1.4027	0.0233	-0.1081	28.0910	0.9477	0.9389	0.80	0.93	1.01	0.95	0.96	4.3
auto	1.0512	1.0581	1.2431	6.1111	3.7224	0.5143	-0.7523	28.2299	0.8604	0.8511	0.46	0.69	0.76	0.65	0.74	3.8
power	0.6270	1.0721	0.2903	3.3211	1.4198	0.1534	-0.0610	27.9740	0.9801	0.9702	0.68	0.85	0.93	0.88	0.90	4.1
entropy	1.0284	3.9810	2.7629	11.0342	7.0204	0.0275	0.0895	28.0001	0.8023	0.7650	0.11	0.27	0.42	0.33	0.39	3.4
attention	0.1699	0.0664	1.2908	6.2123	3.9011	0.5193	-0.7620	28.1530	0.8790	0.8685	0.52	0.75	0.85	0.78	0.80	3.8

TABLE II. EVALUATING THE PERFORMANCE OF POOLING METHODS ON THE MUSIC GENRE CLASSIFICATION TASK USING EVALUATION METRICS TO IDENTIFY WHICH FEATURES EACH POOLING METHOD PRESERVES. THE VALUES OF THE POOLED SPECTROGRAM TO THE ORIGINAL SPECTROGRAM RATIOS OR SIMILARITY SCORES ARE GIVEN IN THE TABLE

Pooling Method	Varia. Preser.	Mean Preser.	Skew.	Kurt.	Entropy	Spars. Level	Gini Coeff.	KL Diver.	Cosi. Simil.	Corr. Coef.	Lips. Cont.	Temp. Var.	Spec. Cent.	Spec. Cont.	Spec. Band.	Comp. Ratio
average	0.8285	0.9988	0.1093	3.0050	1.2711	0.7190	-0.0812	30.0122	0.9922	0.9854	0.67	0.89	0.94	0.90	0.91	4.2
max	1.0775	1.1451	0.4782	4.2239	1.3920	0.0000	-0.1233	30.0988	0.9763	0.9651	0.92	1.00	1.13	1.20	1.12	4.4
min	0.7212	0.8711	-0.4147	2.4568	1.3123	0.9087	-0.0593	30.0775	0.9491	0.9400	0.39	0.60	0.78	0.66	0.71	4.1
mixed	0.9068	0.9394	0.1921	3.1791	1.2922	0.0000	-0.0952	30.0221	0.9655	0.9550	0.74	0.91	0.99	0.92	0.95	4.3
LSP	0.7595	1.0032	0.1582	3.1195	1.3321	0.0000	-0.0744	30.1012	0.9832	0.9728	0.62	0.78	0.92	0.85	0.88	4.2
ESP	0.0118	0.0032	-1.8011	9.8484	4.1311	0.4702	0.8934	30.0440	0.0061	0.0073	0.04	0.14	0.29	0.23	0.27	3.1
LNP	0.7230	-0.9051	-1.5111	5.2457	1.3195	0.0311	0.0955	30.0999	0.9421	0.9304	0.33	0.51	0.68	0.59	0.63	4.1
LSEP	1.0022	0.9199	0.3568	3.4511	1.3855	0.0205	-0.1094	30.0331	0.9703	0.9605	0.78	0.93	1.03	0.94	0.97	4.3
auto	0.5021	0.4777	1.1710	5.7219	3.8110	0.5831	-0.7721	30.1211	0.8891	0.8724	0.48	0.71	0.82	0.75	0.79	3.7
power	0.6872	1.0711	0.2292	3.2744	1.3677	0.1599	-0.0682	30.0114	0.9871	0.9793	0.70	0.84	0.91	0.88	0.90	4.2
entropy	1.0232	3.9822	2.8965	11.7652	7.1333	0.0344	0.0783	30.0552	0.7985	0.7593	0.10	0.26	0.40	0.32	0.37	3.3
attention	0.1799	0.0822	1.1532	5.8799	3.8750	0.5675	-0.7744	30.0901	0.8944	0.8822	0.54	0.73	0.84	0.77	0.81	3.7

TABLE III. EVALUATING THE PERFORMANCE OF POOLING METHODS ON THE ENVIRONMENTAL SOUND CLASSIFICATION TASK USING EVALUATION METRICS TO IDENTIFY WHICH FEATURES EACH POOLING METHOD PRESERVES. THE VALUES OF THE POOLED SPECTROGRAM TO THE ORIGINAL SPECTROGRAM RATIOS OR SIMILARITY SCORES ARE GIVEN IN THE TABLE

Pooling Method	Varia. Preser.	Mean Preser.	Skew.	Kurt.	Entropy	Spars. Level	Gini Coeff.	KL Diver.	Cosi. Simil.	Corr. Coef.	Lips. Cont.	Temp. Var.	Spec. Cent.	Spec. Cont.	Spec. Band.	Comp. Ratio
average	0.7919	0.9962	0.1274	3.0256	1.2915	0.7012	-0.0885	32.6579	0.9876	0.9812	0.62	0.85	0.93	0.81	0.87	4.0
max	1.0477	1.1240	0.4873	4.1285	1.3757	0.0000	-0.1294	32.6133	0.9723	0.9621	0.94	1.00	1.11	1.22	1.09	4.5
min	0.6937	0.8690	-0.3982	2.5127	1.3512	0.9041	-0.0679	32.6632	0.9451	0.9370	0.38	0.59	0.78	0.64	0.73	4.1
mixed	0.8937	0.9326	0.2147	3.1846	1.3033	0.0000	-0.1051	32.6127	0.9584	0.9503	0.71	0.88	0.95	0.87	0.91	4.3
LSP	0.7412	1.0087	0.1623	3.1178	1.3457	0.0000	-0.0837	32.6407	0.9792	0.9734	0.58	0.75	0.88	0.79	0.83	4.2
ESP	0.0001	-0.0009	-1.8345	9.7265	4.1431	0.4610	0.9087	32.6390	0.0021	0.0032	0.02	0.09	0.22	0.18	0.25	3.1
LNP	0.7657	-1.0023	-1.5124	5.1473	1.3103	0.0365	0.0861	32.5663	0.9432	0.9315	0.31	0.48	0.66	0.59	0.64	4.0
LSEP	1.0093	0.9129	0.3719	3.4621	1.3852	0.0186	-0.1176	32.5853	0.9615	0.9574	0.81	0.91	1.02	0.95	0.97	4.4
auto	0.9841	0.9759	1.1872	5.8234	3.8288	0.5749	-0.7837	32.6675	0.8754	0.8622	0.48	0.67	0.74	0.63	0.72	3.6
power	0.6691	1.0643	0.2365	3.2783	1.3886	0.1657	-0.0722	32.5257	0.9845	0.9751	0.66	0.83	0.91	0.86	0.89	4.1
entropy	1.0191	3.9962	2.9174	11.8932	7.1555	0.0389	0.0731	32.6040	0.7893	0.7532	0.09	0.22	0.35	0.27	0.33	3.2
attention	0.1841	0.0759	1.1623	5.9128	3.8288	0.5749	-0.7837	32.6395	0.8907	0.8731	0.53	0.71	0.82	0.75	0.78	3.6

VI. DISCUSSION

By looking at the spectrogram analysis in Section IV-B and results in Section V, we can identify some important points. The relationship between the effectiveness of pooling methods and the type of audio task becomes particularly evident when comparing the results across the above datasets. As we identified in Section IV-B, speech emotion classification [61] is predominantly a local-detail-driven task. Emotional expression in speech is often encoded in short-term fluctuations and nuanced changes in tone, pitch, and duration. Pooling methods that preserve such transient dynamics, particularly max, auto, and LSEP pooling, are better suited to this domain. Music genre classification [62], by contrast, places the highest emphasis on global structure preservation. The ability to maintain continuity in tonal progression, harmonic layering, and rhythmic patterns is vital for distinguishing among musical genres. Here, pooling strategies such as average, auto, and LSEP pooling that prioritize smoothness and structural fidelity outperform those that emphasize localized variance. Environmental sound classification [63] stands out as a domain that requires a balanced approach to both local and global feature retention. The simultaneous need to detect transient events and maintain scene-level continuity necessitates pooling methods that can handle both aspects effectively, such as auto, mixed,

and LSEP pooling.

Therefore, when applying pooling in all types of audio applications, it is important to look at the global structure preservation and local detail retention of the pooling methods. In global structure preservation, the primary focus is to retain the overall structure and temporal patterns of the spectrogram. This is particularly relevant in applications, where the recognition of broad patterns or sequences is critical for identifying or classifying audio signals. Localized Detail Retention focuses on retaining localized and distinctive patterns within the spectrogram that are critical for identifying or classifying audio signals. Unlike global patterns, these features are often small and transient but hold significant importance for understanding the finer details of audio characteristics. Beyond merely retaining these features, it is important to evaluate how transformations affect their clarity and interpretability. Overly aggressive pooling can blur critical details, diminishing the spectrogram's discriminative power for downstream tasks. Conversely, pooling methods that balance dimensionality reduction with effective feature retention can improve computational efficiency without compromising performance. The effectiveness of pooling should therefore be assessed in the context of the specific application. Comparative analyses of pooling methods can shed light on their strengths and limitations, ensuring that the preserved spectrogram features align with the requirements of the intended task. This alignment is essential for optimizing accuracy, performance, and robustness in audio analysis applications.

VII. Conclusion

This study presents a comprehensive evaluation of pooling techniques in audio analysis, focusing on their roles in global structure preservation and localized detail retention within spectrograms. By introducing diverse evaluation metrics across four domains this research uncovered the nuanced strengths and limitations of twelve pooling methods. The findings highlight the critical importance of selecting pooling techniques based on task-specific requirements. Max pooling demonstrates effectiveness in capturing localized features essential for tasks such as emotion recognition and transient event detection. In contrast, average pooling excels at preserving global patterns vital for applications such as music genre classification and acoustic scene analysis. Entropy pooling, with its ability to retain diverse and intricate information, emerges as particularly suitable for complex audio tasks. In order to map the use of pooling methods into audio applications, this study categorizes audio applications into two primary focuses, global structure preservation and localized detail retention based on its characteristics. This categorization provides a practical framework for aligning pooling strategies with specific application needs. Beyond task-specific insights, the study emphasizes the importance of innovative evaluation metrics, such as variance, entropy, sparsity, and similarity measures, in assessing pooling methods. These metrics shift the focus from traditional downstream task accuracy to the ability of pooling techniques to preserve critical spectrogram features. In this study, the most important evaluation metrics are evaluated to identify which metrics are more critical in feature extraction. In the future, the development of more adaptive and hybrid pooling techniques holds promise for achieving an ideal balance between global structure preservation and localized feature retention. Furthermore, integrating pooling methods with advanced neural architectures and self-supervised learning frameworks offers significant potential to advance audio analysis systems.

Acknowledgment

The authors acknowledge the support received from the LK Domain Registry in publishing this study.

References

[1] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," pp. 131–135, 2017.

[2] P. Lopez-Meyer, J. A. del Hoyo Ontiveros, H. Lu, and G. Stemmer, "Efficient end-to-end audio embeddings generation for audio classification on target applications," pp. 601–605, 2021.

[3] V. Passricha and R. K. Aggarwal, "A comparative analysis of pooling strategies for convolutional neural network based hindi asr," *Journal of ambient intelligence and humanized computing*, vol. 11, no. 2, pp. 675–691, 2020.

[4] W. Lin, M.-W. Mak, and L. Yi, "Learning mixture representation for deep speaker embedding using attention." pp. 210–214, 2020.

[5] P. Safari and J. Hernando, "Self multi-head attention for speaker recognition," 2019.

[6] M. Rouvier, P.-M. Bousquet, and J. Duret, "Study on the temporal pooling used in deep neural networks for speaker verification," pp. 501–505, 2021.

[7] Q. Shi, H. Luo, and J. Han, "Subspace pooling based temporal features extraction for audio event recognition." pp. 3850–3854, 2019.

[8] K. Lee, Z. Hyung, and J. Nam, "Acoustic scene classification using sparse feature learning and event-based pooling," pp. 1–4, 2013.

[9] X. Cai, D. Yu, D. Liu, and M. Wu, "Weakly and semi-supervised learning for sound event detection using image pretrained convolutional recurrent neural network, weighted pooling and mean teacher method," vol. 2010, no. 1, p. 012108, 2021.

[10] D. Yu, X. Cai, D. Liu, and Z. Liu, "Semi-supervised sound event detection using multi-scale convolu-tional recurrent neural network and weighted pooling," Tech. Rep., DCASE2021 Challenge, Tech. Rep., 2021.

[11] X. Lu, P. Shen, S. Li, Y. Tsao, and H. Kawai, "Temporal attentive pooling for acoustic event detection." pp. 1354–1357, 2018.

[12] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[13] E. Salah, K. Amine, K. Redouane, and K. Fares, "A fourier transform based audio watermarking algorithm," *Applied Acoustics*, vol. 172, p. 107652, 2021.

[14] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.

[15] S. Atito, M. Awais, W. Wang, M. D. Plumbley, and J. Kittler, "Asit: Local-global audio spectrogram vision transformer for event classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[16] A. Lependin, P. Ladygin, V. Karev, and A. Mansurov, "Fourier chromagrams for fingerprinting, verification and authentication of digital audio recordings," pp. 263–275, 2023.

[17] A. Maccagno, A. Mastropietro, U. Mazziotta, M. Scarpiniti, Y.-C. Lee, and A. Uncini, "A cnn approach for audio classification in construction sites," *Progresses in Artificial Intelligence and Neural Systems*, pp. 371–381, 2021.

[18] J.-W. Hwang, R.-H. Park, and H.-M. Park, "Efficient audio-visual speech enhancement using deep u-net with early fusion of audio and video information and rnn attention blocks," *IEEE Access*, vol. 9, pp. 137 584–137 598, 2021.

[19] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," pp. 646–650, 2022.

[20] A. K. Das and R. Naskar, "A deep learning model for depression detection based on mfcc and cnn generated spectrogram features," *Biomedical Signal Processing and Control*, vol. 90, p. 105898, 2024.

[21] T. Arias-Vergara, P. Klumpp, J. C. Vasquez-Correa, E. Nöth, J. R. Orozco-Arroyave, and M. Schuster, "Multi-channel spectrograms for speech processing applications using deep learning methods," *Pattern Analysis and Applications*, vol. 24, pp. 423–431, 2021.

[22] O. A. Onasoga, N. Yusof, and N. H. Harun, "Audio classification-feature dimensional analysis," pp. 775–788, 2021.

[23] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[24] X. Wang and L. Xu, "Speech perception in noise: Masking and unmasking," *Journal of Otology*, vol. 16, no. 2, pp. 109–119, 2021.

[25] J. Pons, O. Slizovskaia, R. Gong, E. Gomez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," pp. 2744–2748, 2017.

[26] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent systems with applications*, vol. 16, p. 200115, 2022.

[27] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (avef): A deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2019.

[28] J. Pons and X. Serra, "Randomly weighted cnns for (music) audio classification," pp. 336–340, 2019.

[29] W. Li, P. Wang, R. Xiong, and X. Fan, "Spiking tucker fusion transformer for audio-visual zero-shot learning," *IEEE Transactions on Image Processing*, 2024.

[30] P. Satti, N. Sharma, and B. Garg, "Min-max average pooling based filter for impulse noise removal," *IEEE Signal Processing Letters*, vol. 27, pp. 1475–1479, 2020.

[31] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," pp. 464–472, 2016.

[32] Y. Han, S. Lee, J. Nam, and K. Lee, "Sparse feature learning for instrument identification: Effects of sampling and pooling methods," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2290–2298, 2016.

[33] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[34] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," pp. 530–546, 2014.

[35] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," pp. 1713–1721, 2015.

[36] Y. Liu, H. Chen, Y. Wang, and P. Zhang, "Power pooling: An adaptive pooling function for weakly labelled sound event detection," pp. 1–7, 2021.

[37] Y. Liu, C. Chen, J. Kuang, and P. Zhang, "Semi-supervised sound event detection based on mean teacher with power pooling and data augmentation," pp. 4–6, 2020.

[38] C. Nalmpantis, L. Vrysis, D. Vlachava, L. Papageorgiou, and D. Vrakas, "Noise invariant feature pooling for the internet of audio things," *Multimedia Tools and Applications*, vol. 81, no. 22, pp. 32 057–32 072, 2022.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[40] K. Zhang, Z. Wu, J. Jia, H. Meng, and B. Song, "Query-by-example spoken term detection using attentive pooling networks," pp. 1267–1272, 2019.

[41] H. Phan, O. Y. Chén, L. Pham, P. Koch, M. De Vos, I. McLoughlin, and A. Mertins, "Spatio-temporal attention pooling for audio scene classification," 2019.

[42] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," pp. 436–454, 2020.

[43] C. M. Eckhardt, S. J. Madjarova, R. J. Williams, M. Ollivier, J. Karlsson, A. Pareek, and B. U. Nwachukwu, "Unsupervised machine learning methods and emerging applications in healthcare," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 31, no. 2, pp. 376–381, 2023.

[44] S. A. Sepúlveda-Fontaine and J. M. Amigó, "Applications of entropy in data analysis and machine learning: A review," *Entropy*, vol. 26, no. 12, p. 1126, 2024.

[45] L. Tang, H. Tian, H. Huang, S. Shi, and Q. Ji, "A survey of mechanical fault diagnosis based on audio signal analysis," *Measurement*, p. 113294, 2023.

[46] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends in Hearing*, vol. 27, p. 23312165231209913, 2023.

[47] S. Yadav, D. M. Yadav, and K. R. Desai, "A comprehensive survey of automatic dysarthric speech recognition," *Int J Inf & Commun Technol ISSN*, vol. 2252, no. 8776, p. 8776.

[48] A. Natsiou and S. O'Leary, "Audio representations for deep learning in sound synthesis: A review," pp. 1–8, 2021.

[49] Z. Liu, T. Shao, and X. Zhang, "Bcg signal analysis based on improved vmd algorithm," *Measurement*, vol. 231, p. 114631, 2024.

[50] S. Yadav, D. M. Yadav, and K. R. Desai, "A comprehensive survey of automatic dysarthric speech recognition," *Int J Inf & Commun Technol ISSN*, vol. 2252, no. 8776, p. 8776.

[51] S. Tanberk, V. Dağlı, and M. K. Gürkan, "Deep learning for videoconferencing: A brief examination of speech to text and speech synthesis," pp. 506–511, 2021.

[52] S.-W. Park, J.-S. Ko, J.-H. Huh, and J.-C. Kim, "Review on generative adversarial networks: focusing on computer vision and its applications," *Electronics*, vol. 10, no. 10, p. 1216, 2021.

[53] T. Rathi and M. Tripathy, "Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review," *Speech Communication*, p. 103102, 2024.

[54] A. Klapuri and M. Davy, "Signal processing methods for music transcription," 2007.

[55] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," vol. 1, pp. 113–116, 2002.

[56] I. Bozhilov, R. Petkova, K. Tonchev, and A. Manolova, "A systematic survey into compression algorithms for three-dimensional content," *IEEE Access*, 2024.

[57] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[58] B. L. Sturm, "The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[59] J. J. Bosch, F. Fuhrmann, and P. Herrera, "Irmas: a dataset for instrument recognition in musical audio signals," 2014.

[60] K. J. Piczak, "Esc: Dataset for environmental sound classification," pp. 1015–1018, 2015.

[61] G. Mohmad and R. Delhibabu, "Speech databases, speech features and classifiers in speech emotion recognition: A review," *IEEE Access*, 2024.

[62] V. Lyberatos, S. Kantarelis, E. Dervakos, and G. Stamou, "Challenges and perspectives in interpretable music auto-tagging using perceptual features," *IEEE Access*, 2025.

[63] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent systems with applications*, vol. 16, p. 200115, 2022.