SOUND CLASSIFICATION AND USING SPECTROGRAM IMAGE

[1] K.PATHAN, [2]D.ANITHA, [3]V.AMRUTHA, [4]B.KALYANI, [5]K.MARY, [6]E.SUCHITHRA

[1] ASSOCIATE PROFESSOR, DEPT OF ECE, Dr.SAMUEL GEORGE INSTITUTE OF ENGINEERING AND TECHNOLOGY, MARKAPUR

[2,3,4,5,6]U.G STUDENT, DEPT OF ECE, Dr.SAMUEL GEORGE INSTITUTE OF ENGINEERING AND TECHNOLOGY, MARKAPUR

ABSTRACT

Sound classification using spectrogram images is a cutting-edge approach that leverages image processing techniques to analyze and categorize audio signals. In this method, raw audio data is converted into spectrograms—graphical representations of sound frequencies over time. These spectrograms capture essential acoustic features, making them highly suitable for processing using computer vision and deep learning models.

The proposed system integrates image processing techniques with machine learning algorithms, particularly Convolutional Neural Networks (CNNs), to classify different types of sounds, including speech, music, environmental noises, and industrial sounds. By transforming audio classification into an image-based problem, the system benefits from robust feature extraction and classification capabilities commonly used in computer vision. This approach enhances classification accuracy, improves computational efficiency, and enables real-time sound recognition.

This technique has widespread applications in fields such as speech recognition, security surveillance, wildlife monitoring, and healthcare diagnostics. Experimental evaluations demonstrate the system’s high classification accuracy, validating its effectiveness for real-world sound analysis tasks.

INTRODUCTION

With applications ranging from environmental sound analysis and speech recognition to security and healthcare, sound categorisation is a quickly expanding area of artificial intelligence (AI) and signal processing. Mel-frequency cepstral coefficients (MFCCs) and other extracted acoustic properties, as well as raw audio waveforms, are the foundation of traditional sound classification methods. Nevertheless, a different and very successful method uses image processing techniques for categorisation after transforming audio information into visual representations, like spectrogram images. An illustration of the frequency content of a sound wave over time is called a spectrogram. The audio stream is broken down into its frequency components using a Mel spectrogram analysis or Short-Time Fourier Transform (STFT) to create it. Through this transition, the sound's patterns and features can be represented as images, enabling the use of potent deep learning and computer vision algorithms for categorisation. Researchers can use pre-trained convolutional neural networks (CNNs) and other sophisticated image processing methods to increase classification accuracy by approaching sound as an image-based problem. There are various benefits of using spectrogram images for sound classification. Both the temporal and spectral aspects of sound are captured, enabling efficient feature extraction. Additionally, complex patterns in spectrograms can be recognised by deep learning models trained on images, like CNNs, which results in reliable and effective categorisation. Applications including speech emotion recognition, music genre categorisation, environmental sound recognition, and medical diagnostics (e.g., cough analysis for respiratory disease detection) are where this technology excels.

EXISTING METHOD

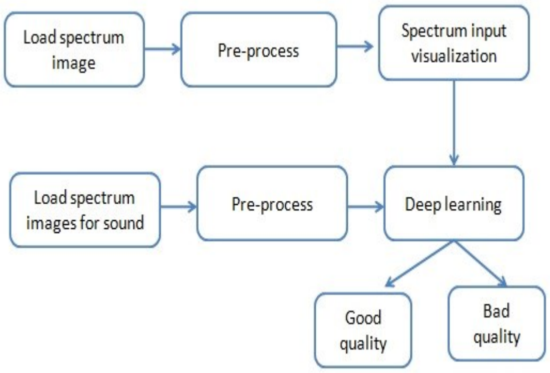
Medical diagnostics, ambient sound monitoring, speech recognition, and many other domains depend heavily on sound classification. Conventional methods mostly involved the analysis of unprocessed audio waveforms and the extraction of manually created characteristics. But new developments in image processing have made it possible for deep learning models to obtain improved classification accuracy by converting audio inputs into spectrogram visuals. A popular technique for classifying sounds is to turn the audio waveform into a spectrogram, which shows the frequency content with time graphically. By breaking down the signal into time-frequency components, the Short-Time Fourier Transform (STFT) is the most often used transformation method. Mel-Frequency Cepstral Coefficients (MFCCs) and Mel Spectrograms are also often used techniques, especially in the classification of music and speech. Once spectrograms are generated, various image processing and deep learning techniques are applied to classify audio signals. Traditional methods relied on feature extraction techniques such as edge detection, texture analysis, and color mapping to identify distinctive patterns in spectrogram images. However, modern advancements leverage convolutional neural networks (CNNs), which automatically learn hierarchical features, resulting in significantly improved classification accuracy. CNN architectures such as VGG16, ResNet, and EfficientNet have been optimized to process spectrograms as image inputs, further enhancing performance.

Another effective approach involves incorporating recurrent neural networks (RNNs) or hybrid CNN-RNN models to analyze spectrogram sequences, capturing the temporal dependencies in sound. Techniques such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) are frequently combined with CNNs to refine classification accuracy. This hybrid approach is particularly beneficial for tasks where the temporal characteristics of audio are crucial, such as speech emotion recognition and music genre classification.

PROPOSED METHOD

Classifying sounds is an essential task in many domains, such as security applications, environmental monitoring, and speech recognition. Conventional techniques use machine learning models and signal processing that have been learnt on unprocessed audio waveforms. On the other hand, converting audio signals into spectrogram images and using image processing algorithms for categorisation is a successful alternative method. Our suggested technique makes use of this strategy to improve sound categorisation tasks' accuracy and efficiency. Gathering raw audio signals from various sources, like microphones or previously recorded datasets, is the first step in the process. These audio signals undergo preprocessing in order to eliminate noise and standardise parameters like duration and sampling rate. Following cleaning, the audio signals are converted into spectrogram images using methods such as Mel-Frequency Cepstral Coefficients (MFCCs) or Short-Time Fourier Transform (STFT). It is simpler to recognise distinctive patterns connected to various sound classes when using spectrograms, which show the frequency components of a sound across time. Following generation, the spectrogram images are processed using image processing techniques to improve feature extraction. Filtering to draw attention to pertinent frequency characteristics, edge identification, and contrast enhancement are common preprocessing methods. Convolutional Neural Networks (CNNs) with a deep learning foundation are then used to categorise the spectrogram images. Using labelled spectrogram datasets, the CNN model is trained to identify patterns linked to several sound categories, including music, voice, alarms, and ambient noises.

SYSTEM DESIGN



DESCRIPTION OF PROPOSED WORK

1. Loading the Spectrum Image

The first step involves converting an audio file into a spectrogram image, a visual representation of sound frequencies over time. This is achieved by loading the audio file (e.g., .wav, .mp3) using libraries like Librosa or pydub. The audio is then transformed using techniques such as Short-Time Fourier Transform (STFT) or Mel-frequency cepstral coefficients (MFCC), producing a 2D image where the x-axis represents time, the y-axis represents frequency, and the color intensity represents amplitude (loudness). **Description:** Converts the audio signal into a spectrogram, making it suitable for image processing and deep learning models.

2. Preprocessing the Spectrum Image

Preprocessing ensures the spectrogram images are clear and standardized for deep learning. Common steps include:

- **Normalization:** Rescaling pixel values to a uniform range (e.g., 0 to

1) for improved model performance.

- **Logarithmic Scaling:** Compressing the dynamic range to make quieter sounds more distinguishable.
- **Noise Reduction:** Using techniques like spectral gating or high-pass filtering to remove background noise.
- **Resizing:** Adjusting the image size (e.g., 224×224 pixels) to fit neural network input requirements. **Description:** Enhances spectrogram quality by standardizing size, reducing noise, and optimizing contrast for deep learning.

3. Spectrum Visualization

Visualizing the spectrogram helps understand sound characteristics and patterns. Using libraries like Matplotlib or Seaborn, the spectrogram can be plotted to analyze frequency variations and intensity spikes.
Description: Provides insights into sound behavior, helping identify key patterns useful for model training.

4. Loading Spectrograms for Classification

A dataset of labeled spectrogram images is prepared, representing various sound categories such as speech, music, and environmental noises. These images are typically stored in directories corresponding to different sound types and loaded using image processing libraries like PIL or OpenCV. **Description:** Organizes and labels spectrogram images for supervised learning and model evaluation.

5. Preprocessing Spectrogram Images

This step further refines spectrograms before model training by applying:

- **Resizing:** Ensuring all images have a consistent dimension (e.g., 224×224 pixels).
- **Data Augmentation:** Applying transformations such as rotation, flipping, and scaling to expand the dataset and improve generalization.
- **Grayscale Conversion:** Simplifying images to grayscale when color information is unnecessary, reducing computational load. **Description:** Standardizes spectrogram images and applies enhancements to improve model robustness and classification accuracy.

6. Deep Learning Model Training

A deep learning model, typically a Convolutional Neural Network (CNN), is trained to classify spectrogram images. CNNs effectively extract frequency-based and time-based patterns from spectrograms. The architecture consists of convolutional layers, pooling layers, and fully connected layers that detect key sound features. The training process involves using a loss function (e.g., cross-entropy loss) and an optimizer (e.g., Adam) to improve classification accuracy. **Description:** The CNN model learns sound classification by identifying distinct spectral patterns from spectrogram images.

7. Characteristics of High-Quality Spectrograms

Well-structured spectrograms enhance model performance. High-quality spectrograms exhibit:

- High contrast between frequency bands.
- Minimal noise and artifacts for clearer feature extraction.
- Distinct patterns that deep learning models can easily interpret. **Description:** High-quality spectrograms improve classification accuracy by providing clear, well-defined sound representations.

8. Challenges of Low-Quality Spectrograms

Poorly constructed spectrograms negatively impact classification performance. Common issues include:

- Excessive noise, leading to misleading features.
- Low resolution, causing blurry or incomplete representations.
- Overexposure or underexposure, reducing visibility of key details.
- Artifacts from preprocessing errors, distorting sound patterns. **Description:** Low-quality spectrograms hinder model learning and increase misclassification risks.

FUTURE SCOPE

The integration of spectrogram-based sound classification and image processing presents exciting opportunities for future advancements across various domains. One major area of development is the enhancement of deep learning models to achieve greater accuracy in sound classification. As machine learning continues to progress, advanced convolutional neural networks (CNNs) and transformer-based architectures can be further optimized to recognize intricate sound patterns with higher precision. Additionally, self-supervised learning techniques can be explored to minimize reliance on labeled datasets, allowing models to learn directly from vast amounts of raw audio data.

Another promising direction is the real-time deployment of sound classification on edge devices. With the widespread adoption of Internet of Things (IoT) technologies, spectrogram-based classification can be integrated into embedded systems for applications such as environmental monitoring, wildlife conservation, and security surveillance. Optimized, lightweight deep learning models can be developed for low-power microcontrollers, mobile devices, and smart sensors, enabling efficient and real-time sound recognition in resource-constrained environments.

ADVANTAGES

1. Enhanced Feature Extraction
2. Compatibility with Image-Based Models
3. Robust to Noise Variations
4. Improved Generalization
5. Easy Data Augmentation
6. Visualization & Interpretability

DISADVANTAGES

1. **Computational Complexity**
2. **Loss of Raw Signal Information**
3. **Dependence on Preprocessing Parameters**
4. **Limited Effectiveness for Short Sounds**
5. **Noise Sensitivity in Low-Quality Data**
6. **Memory Intensive**

CONCLUSION

Spectrogram-based sound classification is a powerful technique for accurately identifying and differentiating various audio signals. By converting raw audio waveforms into spectrograms—visual representations of frequency distribution over time—this method enables the use of advanced image processing and deep learning techniques to analyze sound patterns effectively. This approach allows for precise classification of diverse audio events, including speech, music, environmental sounds, and anomalies.

Integrating spectrograms with machine learning or deep learning models, particularly Convolutional Neural Networks (CNNs), significantly improves classification accuracy by capturing intricate spectral and temporal features. Unlike traditional audio feature extraction methods, spectrogram-based classification provides a more intuitive and interpretable representation of sound patterns. This makes it highly suitable for applications such as security surveillance, healthcare monitoring, wildlife sound analysis, and industrial fault detection.

In summary, the use of spectrogram pictures to sound classification bridges the gap between computer vision and audio signal processing by providing a strong and adaptable framework for identifying and classifying sound events. With ongoing developments in signal processing and artificial intelligence, this strategy will be essential to creating intelligent audio-based systems for a range of practical uses.

REFERENCES

1.E. Wold, T. Blum, D. Keislar and J. Wheaten, "Content-based classification search and retrieval of audio", *IEEE Multimedia*, vol. 3, pp. 27-36, Jun. 1996.

2.F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 337-340, May 2011.

3.M. V. Ghiurcau, C. Rusu, R. C. Bilcu and J. Astola, "Audio based solutions for detecting intruders in wild areas", *Signal Process.*, vol. 92, no. 3, pp. 829-840, 2012.

4.A. Rabaoui, M. Davy, S. Rossignol and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance", *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 763-775, Dec. 2008.

5.S. Chu, S. Narayanan and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features", *IEEE Trans. Audio Speech Language Process.*, vol. 17, no. 6, pp. 1142-1158, Aug. 2009.

6.*Sound Classification*, 2017, [online] Available: http://www.paroc.com/knowhow/sound/sound-classification.

7.R. A. Altes, "Detection estimation and classification with spectrograms", *J. Acoust. Soc. Amer.*, vol. 67, no. 4, pp. 1232-1246, 1980.