Identifying ASMR-Style Audio: Development of a Predictive Classification Model

university of groningen

campus fryslân

Thesis

By

Xiaoling Lin

rijksuniversiteit groningen / campus fryslân

University of Groningen - Campus Fryslân

Identifying ASMR-Style Audio: Development of a Predictive Classification Model

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Joshua Schäuble (Voice Technology, University of Groningen)
with second reader **Dr Shekhar Nayak**

Xiaoling Lin (S5476399)

July 11, 2024

Acknowledgements

I would like to express my heartfelt appreciation to all those who have stood by me during the process of completing this thesis.

Above all, my heartfelt thanks go to Matt, for providing suggestions in direction and detailed planning during the initial stages of my thesis conception. Your guidance helped transform a vague idea into a feasible research topic.

I would also like to thank Joshua for his invaluable assistance with the writing process. Your insights and feedback were crucial to the development of this thesis.

Lastly, a special mention to my cat, Porro Jow, who kept me company behind the glow of the computer screen, providing support through the long stretches of writing and research.

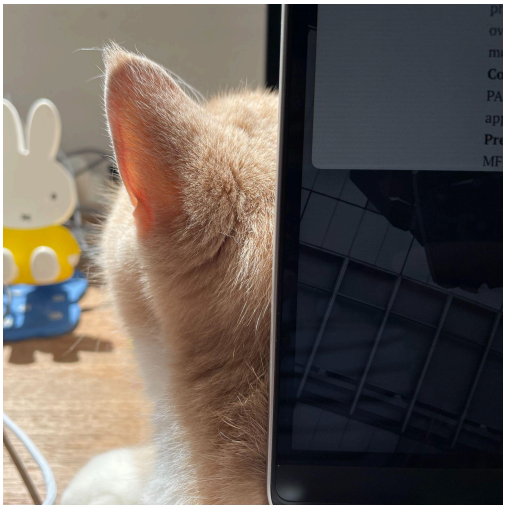

Table of Contents

Abstract

Over the past few years, ASMR (Autonomous Sensory Meridian Response) videos have quickly become a popular genre with a significant emphasis on the auditory aspect in eliciting specific sensory reactions in the audience. However, "ASMR-style audio" lacks a clear definition when compared to common audio. This study aims to fill this gap by creating a model that can predictively differentiate ASMR-style audio from other audio.

The CNN model designed in this study aims to accurately distinguish ASMR-style audio from common audio. The performance of the model is evaluated using accuracy in identifying ASMR-style audio.

This study's findings suggest that synthesizing ASMR-style audio in the future could become possible, allowing individuals to select their preferred ASMR content. By automating the classification of ASMR-style audio, this research not only enhances content curation on streaming platforms but also contributes to the broader field of audio classification and voice technology.

Index Terms - ASMR, audio classification, CNN, voice technology.

1 Introduction

The curious sensation known as Autonomous Sensory Meridian Response (ASMR) has gained significant attention in recent years, among both the general public and the scientific community. ASMR is characterized by a tingling sensation that typically begins on the scalp and moves down the back of the neck and upper spine, often triggered by specific auditory and visual stimuli. Common ASMR triggers include whispering, tapping, and gentle scratching sounds. However, scientific inquiries into the auditory characteristics and overall nature of such phenomenon are still in the early phases of development.

Current research on ASMR primarily focuses on its psychological and physiological effects, demonstrating that ASMR can induce relaxation and reduce heart rate, among other benefits (Poerio et al., 2018). However, the acoustic characteristics that define ASMR triggers are less well understood, and there is a notable gap in the literature regarding the development of automated methods to identify ASMR-style audio from common audio. This study aims to fill this gap by exploring whether a machine learning model can be developed to classify ASMR-style audio.

The importance of this research lies in its potential applications. Accurate identification of ASMR-style audio can enhance content curation on streaming platforms, improve the personalization of ASMR content for therapeutic uses, and contribute to the broader field of audio classification in voice technology. This study also could provide insights into the acoustic properties that make certain sounds soothing or stimulating, furthering our understanding of human auditory perception.

The study aims to achieve two goals: first, to develop an audio classification application for ASMR-style audio using Convolutional Neural Network (CNN); second, to evaluate the model's performance in accurately classifying ASMR-style audio. The rationale behind this research is rooted in the need for a systematic and automated approach to categorize ASMR content, which has been predominantly manual and subjective to date.

This study's contributions to the field of voice technology are multifaceted. By advancing the methodology for ASMR-style audio classification, this research will provide a framework for future ASMR studies to build upon. Furthermore, it will offer practical tools for content creators and platforms to enhance user experience through better content recommendation and personalization.

The methodology involves collecting a diverse dataset of ASMR and non-ASMR-style audio samples and training various machine learning models to determine the most effective approach for classification. The evaluation will focus on the accuracy of identifying ASMR-style audio samples from non-ASMR-style audio samples.

The structure of this paper is as follows: the next section provides a detailed review of the literature on ASMR and audio classification, highlighting key studies and identifying gaps. This is followed by the methodology section, which outlines the data collection, feature extraction, and model training processes. The results section presents the findings from the model evaluations, and the discussion interprets these results in the context of the existing literature. Finally, the conclusion summarizes the key contributions of the study and suggests directions for future research.

This study seeks to advance the field of ASMR research by developing a predictive classification model that adopts machine learning techniques to distinguish ASMR-style audio from common audio. The aim is to enhance the understanding and application of ASMR in various domains, contributing to the broader field of voice technology.

2 Literature Review

This chapter offers a literature review on ASMR and audio classification. It begins with an introduction to ASMR, defining its phenomenology, popularity, and applications, alongside an overview of audio classification techniques. Following this, the focus shifts to deep learning models for audio classification, and their applications in audio tasks. Next, feature extraction and data preprocessing are discussed, detailing techniques like MFCCs and spectrograms, data augmentation, normalization strategies, and handling ASMR-specific features. The discussion then moves to the implementation and evaluation of ASMR-style audio classification models, including model building, training, and evaluation metrics, with a comparative analysis of different models. Finally, the applications and implications of ASMR-style audio classification are explored, covering therapeutic and commercial uses, mental health benefits, and future prospects.

2.1 Introduction to ASMR and Audio Classification

ASMR, or Autonomous Sensory Meridian Response, is a term used to describe a tingling sensation that typically begins on the scalp and moves down the back of the neck and upper spine. This sensation is often triggered by specific auditory and visual stimuli, such as whispering, tapping, and slow hand movements. As Pérez Zarazaga et al. note, "Whisper is also associated with situations of intimacy and speakers may use it to elicit emotion and relaxation. The latter effect contributes to e.g. the popularity of the so-called autonomous sensory meridian response (ASMR) genre on streaming platforms" (Pérez Zarazaga et al., 2023). The phenomenology of ASMR involves a combination of sensory experiences that are often described as calming and pleasurable, leading to its growing popularity.

ASMR has gained significant attention due to its potential therapeutic benefits. As Poerio et al. explain, "ASMR is a reliable and physiologically-rooted experience that may have therapeutic benefits for mental and physical health" (Poerio et al., 2018). This has led to a surge in the production and consumption of ASMR content on various digital platforms, where creators

produce videos designed to evoke these sensations. The growing popularity of ASMR is not only a cultural phenomenon but also an area of academic interest, particularly in understanding its applications and the underlying mechanisms that make certain sounds effective triggers.

The task of identifying and classifying ASMR-style audio presents unique challenges due to the subtle and specific nature of the sounds involved. Traditional audio classification techniques, such as those used in music information retrieval (MIR) and other audio recognition tasks, provide a foundation for this endeavor. Dieleman and Schrauwen describe a common approach in MIR, "Researchers have traditionally relied on a two-stage approach to solve content-based MIR tasks: features are extracted from music audio signals, and are then used as input to a regressor or classifier" (Dieleman & Schrauwen, 2014).

Effective feature extraction is crucial for the success of any audio classification model. Techniques such as Mel-frequency cepstral coefficients (MFCCs) and spectrograms are commonly used to transform raw audio signals into more manageable forms. Piczak describes an approach involving segmented spectrograms, "A deep model consisting of 2 convolutional layers with max-pooling and 2 fully connected layers is trained on a low-level representation of audio data (segmented spectrograms) with deltas" (Piczak, 2015). These features capture important characteristics of the audio signal, facilitating more accurate classification.

Data preprocessing and augmentation also play vital roles in preparing audio data for model training. Normalization strategies ensure that the audio data is consistent and comparable, while data augmentation techniques can help create more robust models by exposing them to a wider variety of training examples. Handling ASMR-style audio features requires careful consideration of the unique properties of these sounds, which often include soft, subtle, and repetitive patterns.

The combination of traditional audio classification techniques with modern deep learning approaches offers a solid framework for identifying ASMR-style audio. Effective feature extraction, data preprocessing, and augmentation are critical for building accurate and reliable models. The therapeutic and commercial applications of ASMR-style audio classification spotlight its importance and substantial influence. By acknowledging the gaps identified and

making use of the progress in audio classification techniques, creating a predictive model for ASMR-style audio can offer fresh perspectives and useful advantages to the field.

2.2 Deep Learning Models for Audio Classification

Deep learning technology has truly transformed the realm of audio classification, providing powerful resources for recognizing and analyzing intricate audio signals, such as those found in ASMR-type recordings. This section delves into various deep learning architectures, including Convolutional Neural Networks (CNNs), Pretrained Audio Neural Networks (PANNs), and the innovative Audio Spectrogram Transformer (AST), and how they have been applied in audio classification tasks.

Convolutional Neural Networks (CNNs) have been at the forefront of audio classification due to their ability to learn hierarchical features from audio data. CNNs are particularly effective in processing and classifying audio signals due to their prowess in capturing spatial hierarchies in data, as Hershey et al. explain, "We use various CNN architectures to classify the soundtracks of a dataset of 70M training videos (5.24 million hours) with 30,871 video-level labels." (Hershey et al., 2017). This capability allows CNNs to handle large and complex datasets, making them suitable for tasks like ASMR-style audio classification, where subtle and nuanced features need to be detected and analyzed.

One of the key advantages of CNNs is their end-to-end learning capability, which minimizes the need for extensive feature engineering. Dieleman and Schrauwen highlight this benefit, "End-to-end learning greatly reduces the need for prior knowledge about the problem, and minimizes the required engineering effort; only the tuning of the model hyperparameters requires some expertise" (Dieleman & Schrauwen, 2014). This streamlined approach is particularly useful in ASMR-style audio classification, where the sounds involved are often subtle and complex, requiring sophisticated models to discern them accurately.

Pretrained Audio Neural Networks (PANNs) represent another significant advancement in audio classification. These models leverage transfer learning to improve performance on specific tasks

by utilizing knowledge gained from large-scale datasets. Kong et al. discuss the efficacy of PANNs, "We propose pretrained audio neural networks (PANNs) trained on the large-scale AudioSet dataset. These PANNs are transferred to other audio related tasks. We investigate the performance and computational complexity of PANNs modeled by a variety of convolutional neural networks" (Kong et al., 2020). PANNs are particularly advantageous for ASMR-style audio classification as they can be fine-tuned for specific tasks, enhancing their accuracy and efficiency.

The innovative Audio Spectrogram Transformer (AST) introduces a purely attention-based model for audio classification, departing from traditional convolutional approaches. Gong, Chung, and Glass describe the AST's unique approach, "The Audio Spectrogram Transformer (AST) is the first convolution-free, purely attention-based model for audio classification, achieving new state-of-the-art results on several benchmarks" (Gong et al., 2021). The AST utilizes attention mechanisms to focus on the most relevant parts of the audio signal, making it highly effective in capturing the intricate patterns characteristic of ASMR-style audio.

Deep learning models have been utilized in various fields for audio classification, showcasing their flexibility and resilience. For instance, Sprengel et al. highlight the application of CNNs in bird species identification, "With novel preprocessing and data augmentation methods, we train a convolutional neural network on the biggest publicly available dataset for bird species identification" (Sprengel et al., 2016). This example highlights the potential of deep learning models to handle diverse and complex audio classification tasks.

In the context of ASMR-style audio classification, distinguishing between phonated and whispered speech is crucial. Pérez Zarazaga et al. emphasize the importance of using appropriate tools, "To process and analyze phonated vs. whispered speech signal, it is important to distinguish between proper tools. Several solutions derived from VAD using deep learning have been studied to separate phonated from whispered speech" (Pérez Zarazaga et al., 2023). Deep learning methods, particularly CNNs and recurrent neural networks (RNNs), have been effective in this area, as Imran et al. note, "Convolutional Neural Networks (CNN) and Recurrent Neural

Networks (RNN) were used to classify audio data, which are models for processing datasets based on deep learning" (Imran et al., 2021).

The integration of attention mechanisms in deep learning models has further enhanced their performance. Wu, Mao, and Zhang illustrate this improvement, "Audio classification using attention-augmented convolutional neural networks shows that the attention mechanism significantly improves the performance of CNNs by focusing on the most relevant parts of the data" (Wu et al., 2018). This focus on relevant data is particularly beneficial for ASMR-style audio classification, where subtle differences in audio signals can significantly impact the classification results.

Transfer learning involves a pre-existing model by making adjustments to address a new but interconnected task, standing as a powerful approach in the field of audio classification. Arora and Haeb-Umbach highlight its use, "Most of the recent works for transfer learning, including the three mentioned above, employ Deep Neural Networks (DNN) as models to transfer the parameters from the source to the target domain" (Arora & Haeb-Umbach, 2017). This technique is particularly useful in ASMR-style audio classification, where large, labeled datasets may not always be available. By transferring knowledge from a related task, models can achieve high performance even with limited labeled data.

In addition to CNNs, PANNs, and ASTs, other deep learning architectures have also been explored for audio classification. For instance, deep belief networks (DBNs) and Long Short-Term Memory (LSTM) networks have shown promise in various audio classification tasks. Lee et al. discuss the training of DBNs, "Convolutional deep belief networks can be efficiently trained using greedy layerwise training, in which the hidden layers are trained one at a time in a bottom-up fashion" (Lee et al., 2009). Similarly, Song et al. highlight the effectiveness of LSTM and CNN models in identifying languages in whispered ASMR speech, achieving high accuracy rates (Song et al., 2023).

The field of audio pattern recognition encompasses several tasks beyond ASMR-style audio classification, including audio tagging, acoustic scene classification, and speech emotion

classification. Kong et al. note the breadth of this field, "Audio pattern recognition is an important research topic in the machine learning area, and includes several tasks such as audio tagging, acoustic scene classification, music classification, speech emotion classification and sound event detection" (Kong et al., 2020). The wide range of uses emphasizes the flexibility of deep learning models in managing a variety of audio classification tasks.

The development of effective preprocessing and data augmentation techniques is crucial for improving model performance. As demonstrated by Piczak, "A deep model consisting of 2 convolutional layers with max-pooling and 2 fully connected layers is trained on a low-level representation of audio data (segmented spectrograms) with deltas" (Piczak, 2015). These techniques help enhance the quality of the input data, leading to more accurate and reliable classification results.

Deep learning models have significantly advanced the field of audio classification, offering powerful tools for identifying and processing complex audio signals, including ASMR-style audio. Convolutional Neural Networks (CNNs), Pretrained Audio Neural Networks (PANNs), and the innovative Audio Spectrogram Transformer (AST) represent some of the most effective architectures for this task. These models have been effectively used in diverse areas, showcasing their adaptability and strength. However, challenges remain, particularly in terms of computational complexity and the need for effective preprocessing and data augmentation techniques. By addressing these challenges, future research can further enhance the performance and applicability of deep learning models for ASMR-style audio classification, contributing new insights and practical benefits to the field.

2.3 Feature Extraction and Data Preprocessing

Feature extraction and data preprocessing are critical steps in developing effective audio classification models, especially for ASMR-style audio. These processes transform raw audio signals into a form that deep learning models can process more efficiently, enhancing their ability to learn relevant patterns. This section details various feature extraction techniques, methods for data augmentation, and strategies for handling ASMR-style audio features.

Feature extraction is fundamental to audio processing as it involves transforming raw audio data into a set of features that can be effectively utilized by machine learning algorithms. Mel Frequency Cepstral Coefficients (MFCCs) and spectrograms are among the most frequently used techniques in this domain. MFCCs are widely used because they approximate the human ear's response more closely than other techniques. Paulin et al. highlight the importance of MFCCs, "Feature extraction is critical for building an efficient ASR system. MFCCs and LPC are among the most commonly used techniques to extract features from the audio signal" (Paulin et al., 2018). MFCCs capture the short-term power spectrum of a sound, providing a compact representation of the audio signal that retains essential information for classification.

Spectrograms, on the other hand, provide a visual representation of the spectrum of frequencies in a signal as it varies with time. Wu, Mao, and Zhang emphasize their utility, "Spectrograms are used to visually represent the spectrum of frequencies of a signal as it varies with time. This transformation is vital for many audio classification tasks" (Wu et al., 2018). By converting time-domain signals into spectrograms, deep learning models can utilize convolutional layers to detect patterns within these visual representations, much like image classification tasks. Lee et al. also support this approach, "We convert time-domain signals into spectrograms for the application of convolutional deep belief networks (CDBNs) to audio data" (Lee et al., 2009).

Additional techniques such as log Mel filterbank features are also effective. Arora and Haeb-Umbach describe their process, "40 log mel filter bank features are extracted and packed in a feature vector for every frame. Each feature vector is normalized by subtracting the sample mean and dividing by the sample standard deviation calculated on the time dimension" (Arora & Haeb-Umbach, 2017). This normalization process ensures that the features have a consistent scale, improving the performance and stability of the learning algorithms.

Preprocessing audio data is essential to enhance the quality and consistency of the dataset. Common preprocessing steps include normalization, Fast Fourier Transformation (FFT), Short-Time Fourier Transformation (STFT), and Mel Filterbank. Imran et al. outline these steps, "Before starting the experiments, the datasets were preprocessed using normalization, Fast Fourier Transformation, Short-Time Fourier Transformation, Mel Filterbank, and Mel Frequency

Cepstral Coefficient" (Imran et al., 2021). Normalization adjusts the amplitude of the audio signals to ensure that all data points have a similar scale, which helps in achieving consistent learning during model training. The Fast Fourier Transformation (FFT) converts time-domain signals into frequency-domain signals, making it easier to analyze the frequency components of the audio. The Short-Time Fourier Transformation (STFT) is a variation of FFT that applies the transformation to short overlapping segments of the signal, providing a time-frequency representation.

The use of Hamming and Hanning windows is common in the computation of spectrograms. Gong, Chung, and Glass describe their methodology, "The input audio waveform is converted into a sequence of 128-dimensional log Mel filterbank features computed with a 25ms Hamming window and 10ms frame shift" (Gong et al., 2021). Similarly, Piczak states, "We find the spectrogram of each audio signal, using a Hanning window of 25ms with a 10ms overlap. This gives us a time-frequency representation, with frequencies plotted along the y-axis and time along the x-axis" (Piczak, 2015).

Data augmentation is a crucial strategy to enhance the training dataset by artificially increasing its size and diversity. This process helps in preventing overfitting and improves the robustness of the model. Paulin et al. define data augmentation, "Data augmentation involves altering the training data using various techniques to increase the size and quality of the dataset" (Paulin et al., 2018). Common data augmentation techniques include adding noise, pitch shifting, time stretching, and random cropping. These methods introduce variability into the training data, making the model more resilient to variations in real-world audio. Kong et al. emphasize the importance of data augmentation and balancing, "Data balancing is a technique used to train neural networks on a highly unbalanced dataset. Data augmentation is a technique used to augment the dataset, to prevent systems from overfitting during training" (Kong et al., 2020).

Label smoothing is another technique applied during training to improve model generalization. Kong et al. describe its application, "We apply label smoothing during training, which replaces the hard zero or one labels with smoothed values" (Kong et al., 2020). This approach helps in making the model less confident in its predictions, thereby reducing the likelihood of overfitting.

ASMR-style audio features are likely to involve subtle and nuanced sounds that are different from typical audio signals. Therefore, specialized techniques are necessary to capture these unique characteristics. For instance, RASTA-PLP features have been effectively used for the detection of whispered speech in noisy environments. Pérez Zarazaga et al. note, "For the detection of whispered speech in noisy environments, RASTA-PLP features have been used successfully" (Pérez Zarazaga et al., 2023).

In addition to RASTA-PLP, Edyson labeling has been found to be highly accurate for annotating ASMR-style audio. Pérez Zarazaga et al. explain, "Labeling with Edyson has been found to be highly accurate. The time needed for labeling does not scale linearly with the amount of audio being annotated" (Pérez Zarazaga et al., 2023). Accurate labeling is essential for training robust models, as it ensures that the training data accurately represents the characteristics of ASMR-style audio.

Feature extraction techniques need to be carefully chosen to capture the subtle nuances of ASMR-style audio. For instance, the use of log Mel filterbank features and spectrograms can help in capturing the frequency variations and temporal dynamics of ASMR-style audio. By incorporating these characteristics along with effective data augmentation strategies, it is possible to greatly improve the model's capacity for identifying and categorizing ASMR-style audio.

Beyond the standard methods, advanced techniques such as deep learning directly on raw waveforms and attention mechanisms offer promising avenues for feature extraction and preprocessing. Aytar, Vondrick, and Torralba suggest an innovative approach, "For the application of deep learning to audio, one approach is to use a deep convolutional network that learns directly from the waveform" (Aytar et al., 2016). This method bypasses traditional feature extraction steps, allowing the model to learn features directly from the raw audio signal.

Attention mechanisms have also been integrated into deep learning models to improve their performance on audio classification tasks. These mechanisms allow the model to focus on the most relevant parts of the audio signal, enhancing its ability to capture important features. Wu,

Mao, and Zhang highlight the benefits of this approach, "Audio classification using attention-augmented convolutional neural networks shows that the attention mechanism significantly improves the performance of CNNs by focusing on the most relevant parts of the data" (Wu et al., 2018).

Machine learning algorithms such as logistic regression, support vector classifiers (SVC), K-nearest neighbors (KNN), random forests, and decision trees have been applied to audio classification tasks with varying degrees of success. Kumar notes, "Machine learning methods have been applied to the classification task to classify music genres. Algorithms used are LogisticRegression, SVC using different kernels (linear, sigmoid, rbf and poly), KNeighborsClassifier, RandomForestClassifier, DecisionTreeClassifier, and GaussianNB" (Kumar, 2023).

Despite the advancements in feature extraction and data preprocessing, several challenges remain. The computational complexity of these techniques can be a barrier to their widespread adoption. Additionally, the variability in ASMR-style audio signals, such as differences in recording quality and environmental noise, poses a challenge for consistent feature extraction and classification.

Feature extraction and data preprocessing are critical components of developing effective ASMR-style audio classification models. Techniques such as MFCCs, spectrograms, and log Mel filterbank features provide valuable representations of audio signals, while data augmentation and normalization enhance the quality and consistency of the training data. Specialized methods for handling ASMR-style features, combined with advanced techniques such as attention mechanisms and deep learning on raw waveforms, offer promising avenues for improving the performance of ASMR-style audio classification models. By addressing the challenges and leveraging these advanced techniques, future research can significantly enhance the accuracy and robustness of ASMR-style audio classification, contributing valuable insights to the field.

2.4 Implementation and Evaluation of ASMR-style audio Classification Models

Building and training models for ASMR-style audio classification involves several steps, from the selection of appropriate model architectures to the implementation of training protocols and the evaluation of model performance using standardized metrics. This section describes the process of constructing and training classification models, explains the evaluation metrics used to assess model performance, and provides a comparative analysis of different models.

To effectively classify ASMR-style audio, it is essential to choose model architectures that can capture the intricate and subtle patterns characteristic of ASMR-style sounds. CNNs have proven to be highly effective for this purpose due to their ability to learn hierarchical feature representations from audio data. Piczak describes the architecture of a typical CNN model used for audio classification, "A deep model consisting of 2 convolutional layers with max-pooling and 2 fully connected layers is trained on a low-level representation of audio data (segmented spectrograms) with deltas" (Piczak, 2015). This structure allows the model to capture both local and global patterns in the audio signal, making it well-suited for ASMR-style audio classification.

In addition to CNNs, Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, are also used in audio classification tasks, especially for sequences where temporal dependencies are crucial. Pérez Zarazaga et al. illustrate the use of RNNs in their work, "We build an RNN classifier to detect the beginning and end of any whispered speech segments, including those embedded in noise, consisting of other acoustic triggers or any other noise" (Pérez Zarazaga et al., 2023). RNNs are particularly useful for tasks involving sequential data, as they can maintain information over time, which is essential for detecting transitions in ASMR-style audio.

When building these models, it is crucial to preprocess the audio data appropriately. This involves extracting relevant features such as Mel Frequency Cepstral Coefficients (MFCCs), which Paulin et al. describe, "The Mel Frequency Cepstral Coefficients (MFCC) features are

calculated for every 25ms window of the audio signal. So, for every audio file, the returned numpy array would be of the shape [Number of frames * 13]" (Paulin et al., 2018). These features provide a compact representation of the audio signal, capturing its essential characteristics while reducing the dimensionality of the input data.

Training deep learning models for ASMR-style audio classification requires large and well-annotated datasets. Hershey et al. discuss the impact of dataset size on model performance, "We train with subsets of YouTube-100M spanning 23K to 70M videos to evaluate the impact of training set size on performance, and we investigate the effects of label set size on generalization by training models with subsets of labels, spanning 400 to 30K" (Hershey et al., 2017). Larger datasets generally lead to better model performance, as they provide more diverse examples for the model to learn from, thereby improving its ability to generalize to unseen data.

Data augmentation techniques are also employed to enhance the training dataset by artificially increasing its size and diversity. These techniques include adding noise, pitch shifting, time stretching, and random cropping. Kong et al. emphasize the importance of data augmentation, "Data augmentation is a technique used to augment the dataset, to prevent systems from overfitting during training" (Kong et al., 2020). By introducing variability into the training data, these methods help the model become more robust to variations in real-world audio.

Assessing audio classification models involves utilizing a variety of metrics to gauge their effectiveness. Typical metrics utilized for this purpose include accuracy, precision, recall, F1-score, and mean average precision (mAP). Accuracy measures the proportion of correctly classified samples, while precision and recall provide insights into the model's ability to correctly identify positive samples and retrieve all relevant samples, respectively. The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both aspects. Arora and Haeb-Umbach describe additional metrics used in their work, "Evaluation of the networks is carried out using two segment-based evaluation metrics: F-measure and acoustic event error rate (AEER), which are calculated as an average of intermediate metrics on 1-second segments" (Arora & Haeb-Umbach, 2017). These metrics provide a comprehensive view of the model's performance across different evaluation criteria.

In the context of ASMR-style audio classification, CNNs have often outperformed other models due to their ability to capture spatial hierarchies in the data. Imran et al. highlight this performance difference, "CNN model outperformed their RNN counterparts and this is not surprising since CNN's have been leading in classification tasks in computer vision" (Imran et al., 2021). The success of CNNs can be attributed to their convolutional layers, which effectively detect patterns and features in the input data, making them highly suitable for audio classification tasks.

However, other models, such as the Audio Spectrogram Transformer (AST), have shown promise in advancing the state-of-the-art in audio classification. Gong, Chung, and Glass describe the superior performance of AST, "The Audio Spectrogram Transformer (AST) outperforms state-of-the-art systems on various audio classification benchmarks, achieving 95.6% accuracy on ESC-50 and 98.1% accuracy on Speech Commands V2" (Gong et al., 2021). The AST leverages attention mechanisms to focus on the most relevant parts of the audio signal, thereby enhancing its classification performance.

In addition to model architecture, the choice of features used for training plays a crucial role in the performance of the model. Lee et al. compare the performance of different feature sets, "The CDBN features consistently outperformed MFCC features when the number of training examples was small" (Lee et al., 2009). This finding highlights the importance of selecting appropriate features that can capture the relevant characteristics of the audio signal, particularly when the training dataset is limited.

The use of attention mechanisms has further improved the performance of deep learning models for audio classification. Wu, Mao, and Zhang illustrate the benefits of this approach, "The attention mechanism significantly improves the performance of CNNs by focusing on the most relevant parts of the data" (Wu et al., 2018). By directing the model's focus to the most informative parts of the audio signal, attention mechanisms enhance the model's ability to detect and classify relevant patterns, making them particularly useful for ASMR-style audio classification.

Combining different approaches can also lead to improved performance. For example, Valente and Hermansky discuss the benefits of combining evidence from different frequency channels, "Combining evidence from different frequency channels, processed in both parallel and hierarchical fashions, improves overall ASR performance, with a reduction in WER by 6.2%" (Valente & Hermansky, 2008). This multi-channel approach leverages complementary information from different parts of the frequency spectrum, enhancing the model's ability to classify audio accurately.

The implementation and evaluation of ASMR-style audio classification models involve selecting appropriate model architectures, preprocessing the audio data, training the models on large and diverse datasets, and evaluating their performance using standardized metrics. CNNs have proven to be highly effective for this task, but other models, such as RNNs and ASTs, also offer promising avenues for further improvement. The use of advanced techniques such as attention mechanisms and data augmentation, along with the careful selection of features, can significantly enhance the performance of these models. Future research can enhance the accuracy and durability of audio classification models by addressing challenges and utilizing advanced techniques, thereby providing beneficial contributions to the field.

2.5 Applications and Implications of ASMR-style audio Classification

The applications and implications of ASMR-style audio classification extend across various domains, notably in therapeutic and commercial sectors, highlighting its significant potential benefits in mental health and media. ASMR, known for its ability to trigger a tingling sensation and induce relaxation, presents substantial opportunities for enhancing well-being and creating engaging content.

The therapeutic potential of ASMR is particularly noteworthy. Research indicates that ASMR can elicit emotional, psychological, and neuro-physiological responses, which can be harnessed for mental health benefits. Pérez Zarazaga et al. highlight the importance of ASMR datasets in human-computer interaction (HCI), "Our large and growing dataset enables whisper-capable, data-driven speech technology and linguistic analysis. It also opens opportunities in e.g. HCI as a

resource that may elicit emotional, psychological and neuro-physiological responses in the listener" (Pérez Zarazaga et al., 2023). This suggests that ASMR can be integrated into therapeutic interventions aimed at reducing stress and anxiety, promoting relaxation, and improving overall mental health.

ASMR-style audio classification can enhance healthcare technologies, particularly in smart home environments designed for elderly care. Arora and Haeb-Umbach note the relevance of acoustic event detection (AED) in healthcare, "AED has also been used immensely in the field of building smart homes with a focus on healthcare for the elderly" (Arora & Haeb-Umbach, 2017). By accurately classifying ASMR-style audio, these systems can be tailored to provide calming sounds in particular situations, which will improve the quality of life for elderly individuals, helping them manage stress and sleep better.

The commercial applications of ASMR are equally compelling. The growing popularity of ASMR content on platforms like YouTube highlights its promising prospects in the realm of media and entertainment. ASMR videos, characterized by whispering, tapping, and various calming sounds, attract millions of viewers seeking relaxation and sensory delight. Developing advanced ASMR-style audio classification models can enhance content recommendation systems, ensuring that users receive personalized suggestions that match their preferences. This can increase user engagement and satisfaction, increasing viewership and revenue for content creators and platforms.

ASMR-style audio classification also has broader implications in various audio recognition tasks. For instance, automatic urban sound classification, as discussed by Salamon, Bello, and Ellinger, can benefit from advances in ASMR-style audio classification, "Automatic urban sound classification can benefit a variety of multimedia applications" (Salamon et al., 2014). The techniques developed for ASMR can be adapted to other domains, improving the accuracy and efficiency of sound classification systems in diverse environments.

Future research in ASMR-style audio classification holds promise for further advancements in this field. Pérez Zarazaga et al. outline potential future work, "Our future work involves using

metadata analysis and machine-assisted human classification to categorize the whisper styles present in the genre and provide labels for further supervised learning" (Pérez Zarazaga et al., 2023). By leveraging metadata and human-assisted classification, researchers can develop more refined and accurate models, enhancing the ability to classify and understand ASMR-style audio.

The applications and implications of ASMR-style audio classification are extensive and varied. From its therapeutic benefits for mental health to its commercial prospects in media and entertainment, this field's advancement can contribute a lot in voice technology. Utilizing advanced deep learning models, transfer learning, and large-scale datasets, researchers can create more precise and reliable ASMR-style audio classification systems, opening up new avenues for improving well-being and producing captivating content.

3 Research Question and Hypothesis

Research Question:

Can a machine learning model be developed to effectively classify ASMR-style audio, including human voices and triggers, from common audio?

Hypothesis:

It is hypothesized that by using machine learning techniques, a classification model can be developed to accurately distinguish ASMR-style audio from common audio. The model's performance will be evaluated based on its accuracy in identifying ASMR-style audio. The model's performance will be robust across different types of ASMR triggers, maintaining effectiveness in distinguishing ASMR-style audio regardless of the specific trigger used.

If the hypotheses are not supported by the results of the experiments and the model does not perform significantly better than baseline models, it may suggest that the acoustic features of ASMR-style audio are not distinct enough for reliable classification, or that additional features and more complex modeling techniques are required to capture the nuances of ASMR-style audio.

4 Methods

This section delivers an in-depth analysis of the data used in both training and assessing the predictive classification model for identifying ASMR-style audio, alongside a detailed breakdown of the research methodology implemented throughout the study. As discussed in the literature review section, Convolutional Neural Networks (CNNs) generally outperform other methods in audio classification tasks. Taking into account the available computing resources and dataset, I have chosen to use a 1D Convolutional Neural Network (Conv1D) for addressing the research question.

4.1 Data

4.1.1 ASMR-Style Audio Dataset

The main dataset used for this study is the ASMR-85 dataset, which was obtained from Kaggle. This dataset was put together by a user named Kanelsnegl and contains a wide variety of ASMR-style audio recordings. It includes different triggers and speech styles typically found in ASMR content. The recordings vary in length, ranging from five minutes to over an hour. To ensure consistent analysis, each audio file was divided into 10-second clips. This decision was made after noticing that the slow-paced speech and inclusion of diverse triggers in ASMR-style content can be effectively represented within 10-second intervals. And a comparative analysis demonstrated that the word count in 10-second ASMR-style clips closely aligns with that of 5-second clips in common audio datasets.

4.1.2 Common Audio Dataset

In contrast to the ASMR dataset, I use the Common Voice dataset from Mozilla. Common Voice is an extensive open-source speech dataset contributed by volunteers globally, featuring a wide variety of accents, ages, and languages. The dataset is thoughtfully organized and verified, incorporating detailed information such as the speaker's age, gender, and accent. In this study, I

exclusively focused on the English subset of the Common Voice dataset, specifically selecting clean speech samples to act as a comparative control group against the ASMR-style audio.

4.1.3 Data Preprocessing

Both datasets were preprocessed to ensure consistency and quality. For the ASMR-85 dataset, the audio was first segmented into 10-second samples. As for the Common Voice dataset, the maintainers had already done the job, providing 5-second clips ready for use. This preprocessing step was crucial to mitigate variations in recording levels, ensuring that the subsequent analysis was based on comparable audio samples.

To further see the differences in audio features between ASMR-style and common voice clips, Mel-Frequency Cepstral Coefficients (MFCCs) were computed for one random sample from both datasets. The following images show the MFCC visualizations of a typical ASMR-style audio clip and a common voice audio clip:

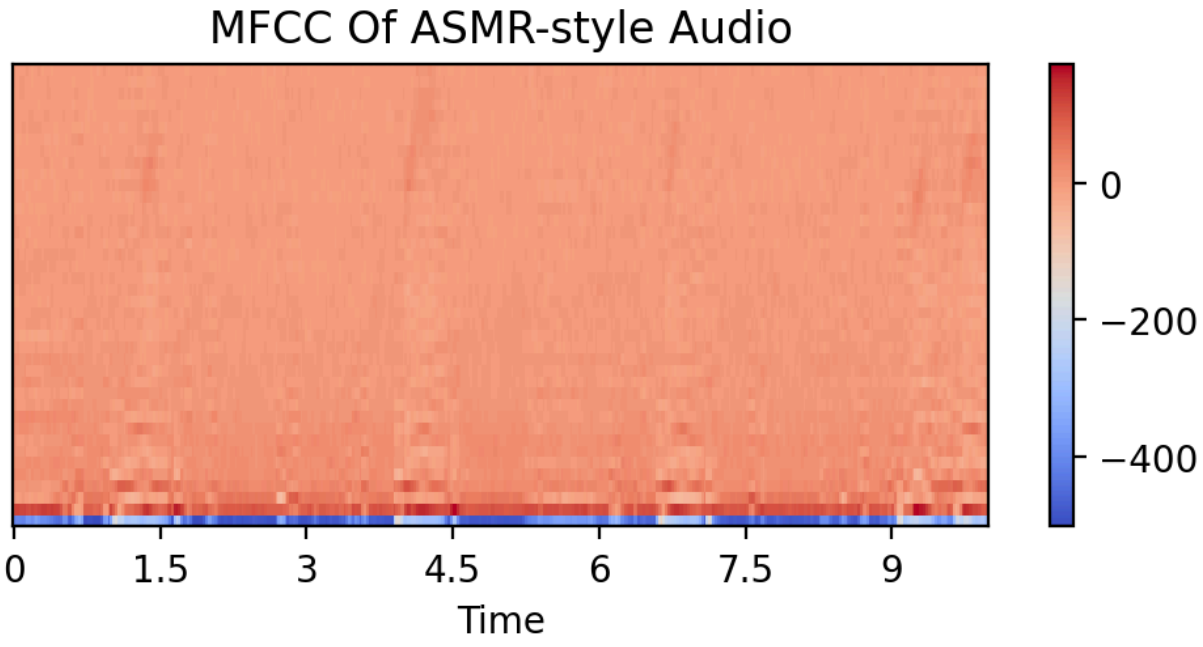

Figure 1: MFCC of an ASMR-style audio (10_clip370.wav)

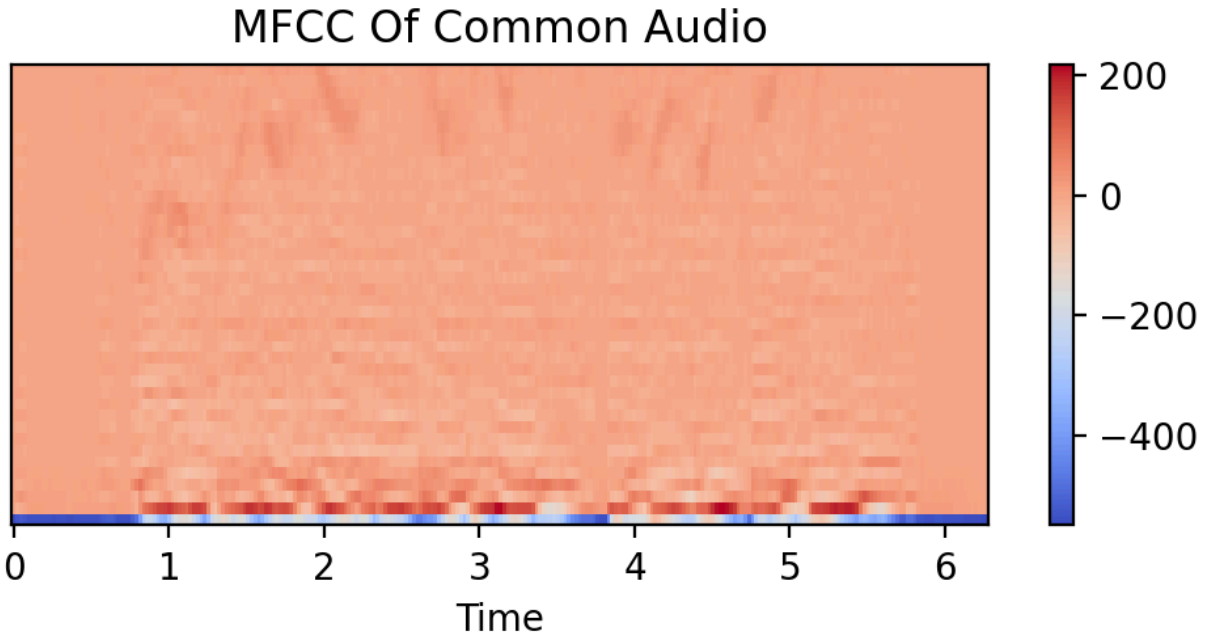

Figure 2: MFCC of a common voice Audio (common_voice_en_19860993.wav)

A chart was generated to demonstrate the distribution of audio samples between the ASMR-style and common audio datasets. This visual representation effectively shows the quantity of samples in each category, allowing for a straightforward comparison:

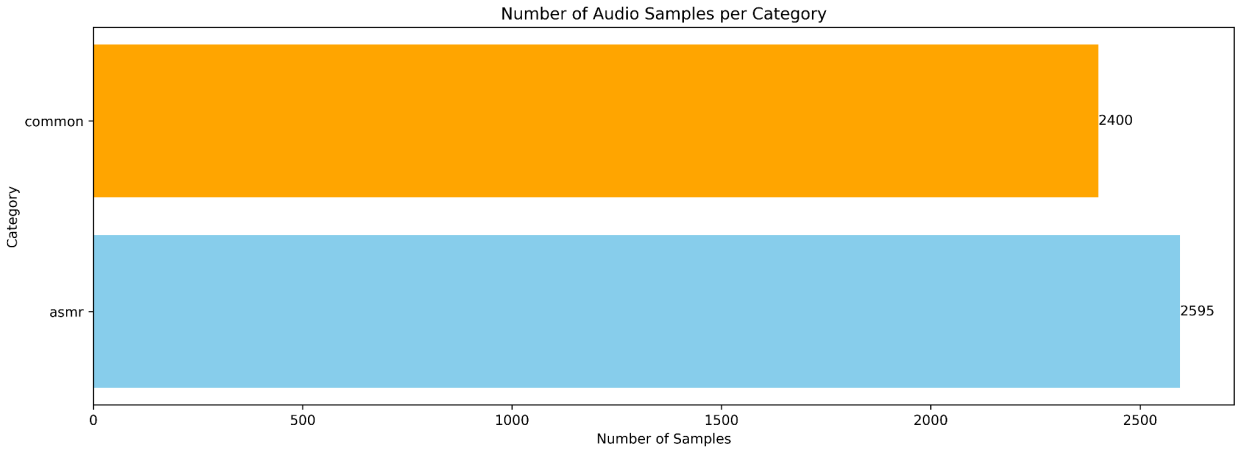

Figure 3: Number of audio samples per category

This chart demonstrates that there are slightly more ASMR-style audio samples compared to common audio samples. This is intentional, as using a little larger amount of “non-standard” training data can often enhance the model's ability to learn distinctive features of the “non-standard” dataset, thereby improving overall training effectiveness.

4.1.4 Ethical Considerations

All datasets used in this study are legally sourced and publicly available, adhering to open data principles. The ASMR-85 dataset and the Common Voice dataset are both licensed for public use, ensuring compliance with copyright regulations. Ethical considerations were also observed by anonymizing any personal identifiers within the datasets, thereby protecting the privacy of the speakers.

4.2 Experiment structure

The experiment was designed to classify ASMR-style audio from common audio using a 1D Convolutional Neural Network (Conv1D). This section outlines the structure of the experiment, detailing the steps taken from data preparation to model evaluation.

4.2.1 Importing the Dataset and Data Preprocessing

The ASMR-85 dataset and Mozilla's Common Voice dataset were obtained from the Kaggle dataset. As mentioned in the previous section, the data was preprocessed for feature extraction.

4.2.2 Feature Extraction and Database Building

To transform audio data into a format compatible with the Conv1D model, I use the librosa library for its feature extraction capabilities. This process includes extracting Mel-Frequency Cepstral Coefficients (MFCC), which detail the frequency distribution across time and allow for the analysis of the audio signals' time and frequency attributes.The process of extracting features involved defining a function to extract features from an audio file, using this function on all audio samples to acquire MFCC features, and then organizing the extracted features into a dataframe with columns for the features and their corresponding class labels.

4.2.3 Building, Training, and Compiling Conv1D Models

The extracted features and class labels were saved into different arrays. Then the data was split into training and testing sets to facilitate the training of the Conv1D model. The architecture of the Conv1D model, summarized in Figure 4, consists of several convolutional layers, each followed by activation functions and pooling layers, culminating in dense layers for final classification.

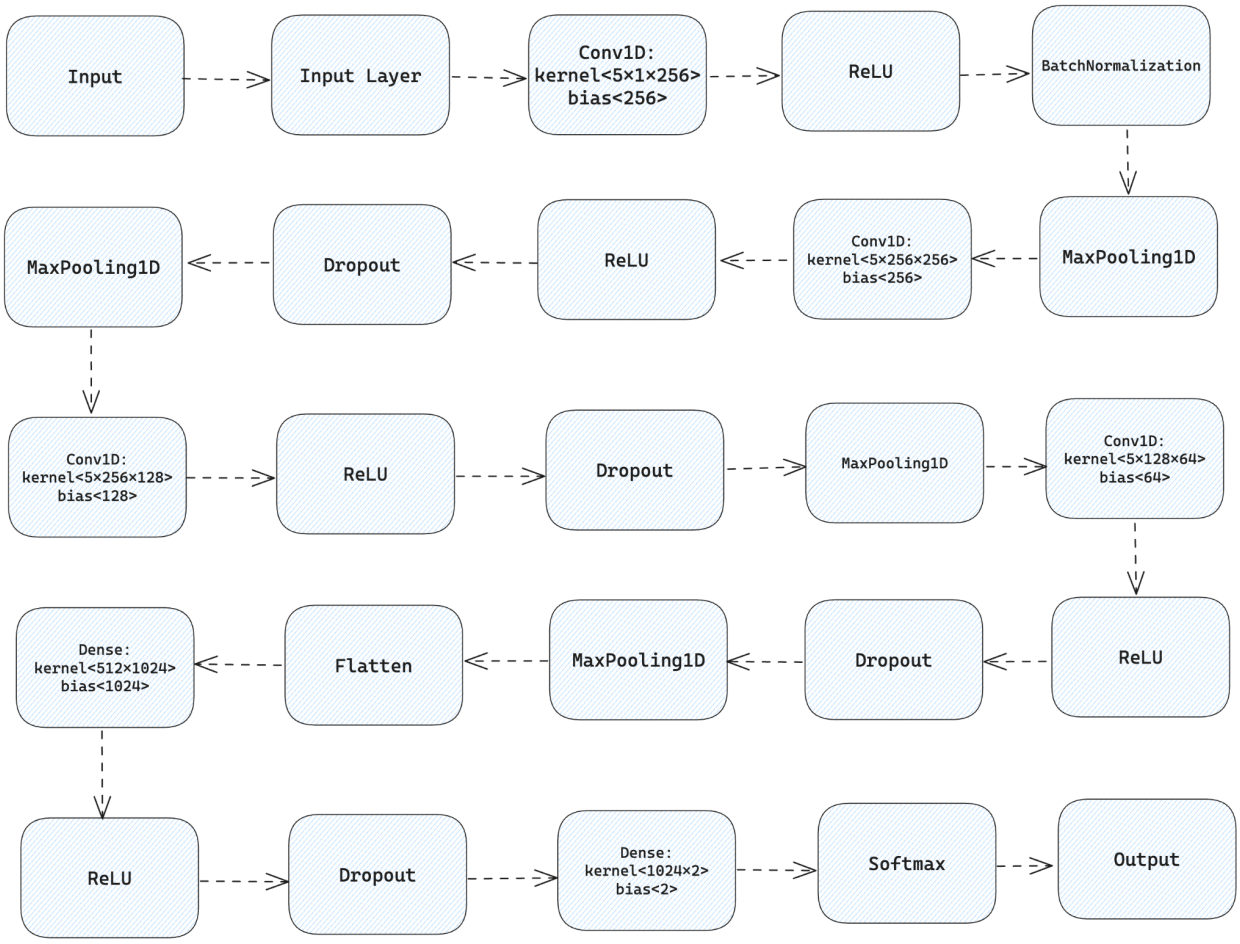

Figure 4: Model Architecture

The network begins with an input layer followed by a series of one-dimensional convolutional layers (Conv1D) with ReLU activation functions and batch normalization layers (BatchNormalization). The network also includes max-pooling layers (MaxPooling1D) and dropout layers (Dropout) to prevent overfitting. The convolutional layers are followed by a

flatten layer (Flatten) that reshapes the data for the dense layers (Dense). The final layer is a dense layer with a softmax activation function for classification. This structure enables the network to efficiently grasp hierarchical features from the provided data.

The model was constructed with reasonable loss functions and optimizers, and it conducted training over numerous epochs with specified batch sizes and learning rates. Throughout the training process, checkpoints were regularly stored to ascertain the most effective model parameters.

Layer (type)	Output Shape	Param #
conv1d_4 (Conv1D)	(None, 128, 256)	1,536
batch_normalization_1 (BatchNormalization)	(None, 128, 256)	1,024
max_pooling1d_4 (MaxPooling1D)	(None, 64, 256)	0
conv1d_5 (Conv1D)	(None, 64, 256)	327,936
dropout_4 (Dropout)	(None, 64, 256)	0
max_pooling1d_5 (MaxPooling1D)	(None, 32, 256)	0
conv1d_6 (Conv1D)	(None, 32, 128)	163,968
dropout_5 (Dropout)	(None, 32, 128)	0
max_pooling1d_6 (MaxPooling1D)	(None, 16, 128)	0
conv1d_7 (Conv1D)	(None, 16, 64)	41,024
dropout_6 (Dropout)	(None, 16, 64)	0
max_pooling1d_7 (MaxPooling1D)	(None, 8, 64)	0
flatten_1 (Flatten)	(None, 512)	0
dense_2 (Dense)	(None, 1024)	525,312
dropout_7 (Dropout)	(None, 1024)	0
dense_3 (Dense)	(None, 2)	2,050

Total params: 1,062,850 (4.05 MB)
Trainable params: 1,062,338 (4.05 MB)
Non-trainable params: 512 (2.00 KB)

Figure 5: Model Parameter Summary

The table (Figure 5) presents a comprehensive overview of the model parameters associated with each layer in the model. It lists the layer type, output shape, and parameter quantity. The output shape specifies the data dimensions at each layer transition. The parameter count includes both trainable and non-trainable parameters, giving insights into the complexity of the model. The

total number of parameters in the network is 1, 062, 850 with 1, 062, 338 trainable parameters and 512 non-trainable parameters.

4.2.4 Predicting the Test Audio

After training, a function was defined to handle prediction tasks. This function extracts features from input audio files, processes them to fit the Conv1D model's input shape, and outputs the predicted class label. The model's performance was evaluated using accuracy in ASMR-style audio classification. The full experiment can be viewed in the Kaggle notebook.

5 Results

The developed Conv1D model aimed to classify ASMR-style audio from common audio. This section shows the model's performance focusing on accuracy and loss metrics over 150 epochs.

The loss per epoch graph (see Figure 6) demonstrates the model's performance in minimizing the error between predicted and actual values during training and validation phases. Initially, both training and validation loss exhibit a steep decline within the first 10 epochs, indicating rapid learning and adjustment of model parameters. After approximately 10 epochs, the loss values for both training and validation datasets converge, remaining low and stable with minor fluctuations. However, from around 80 epochs onward, validation loss begins to show slight increases and fluctuations, while training loss remains consistently low. This divergence suggests potential overfitting, where the model performs slightly better on training data compared to unseen validation data.

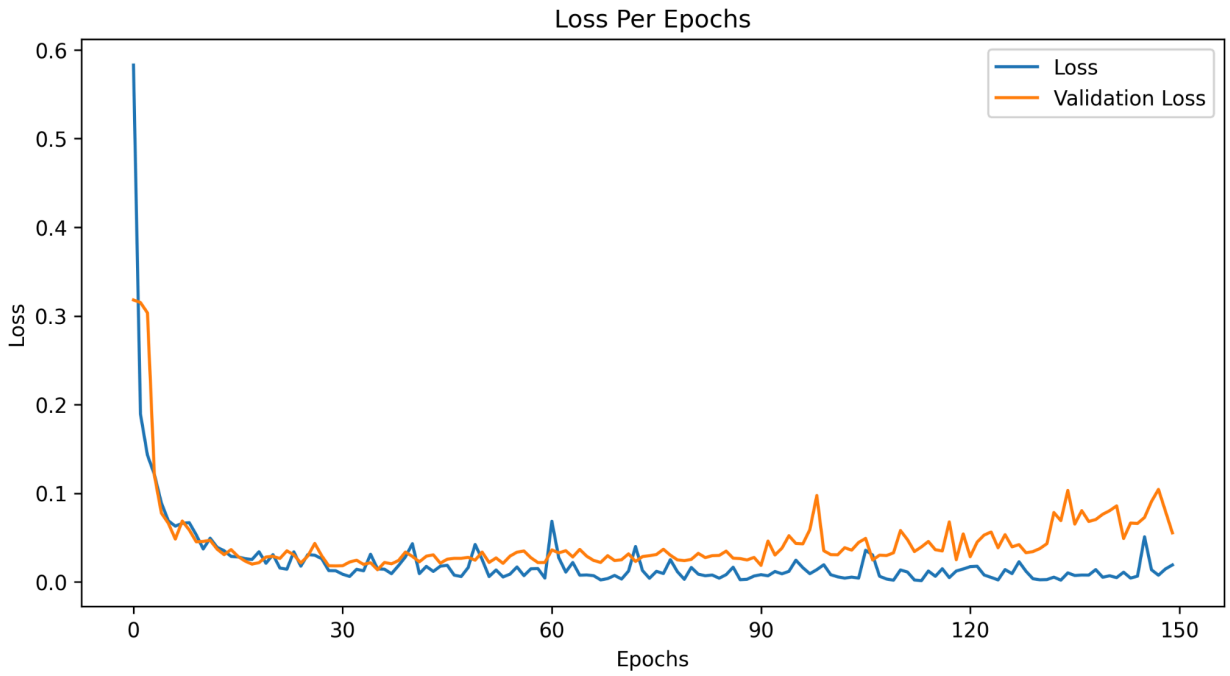

Figure 6: Loss Per Epochs for the Conv1D Model

The accuracy per epoch graph (see Figure 7) presents the accuracy rates for the Conv1D model during training and validation across the same epochs. The model's accuracy for both training and validation data improves significantly within the first 10 epochs, stabilizing around the 98% mark. Post initial learning phase, the accuracy rates for both training and validation remain high and relatively stable, consistently hovering around 98-99%. There is a high degree of consistency between training and validation accuracy, indicating that the model generalizes well to unseen data without significant variance.

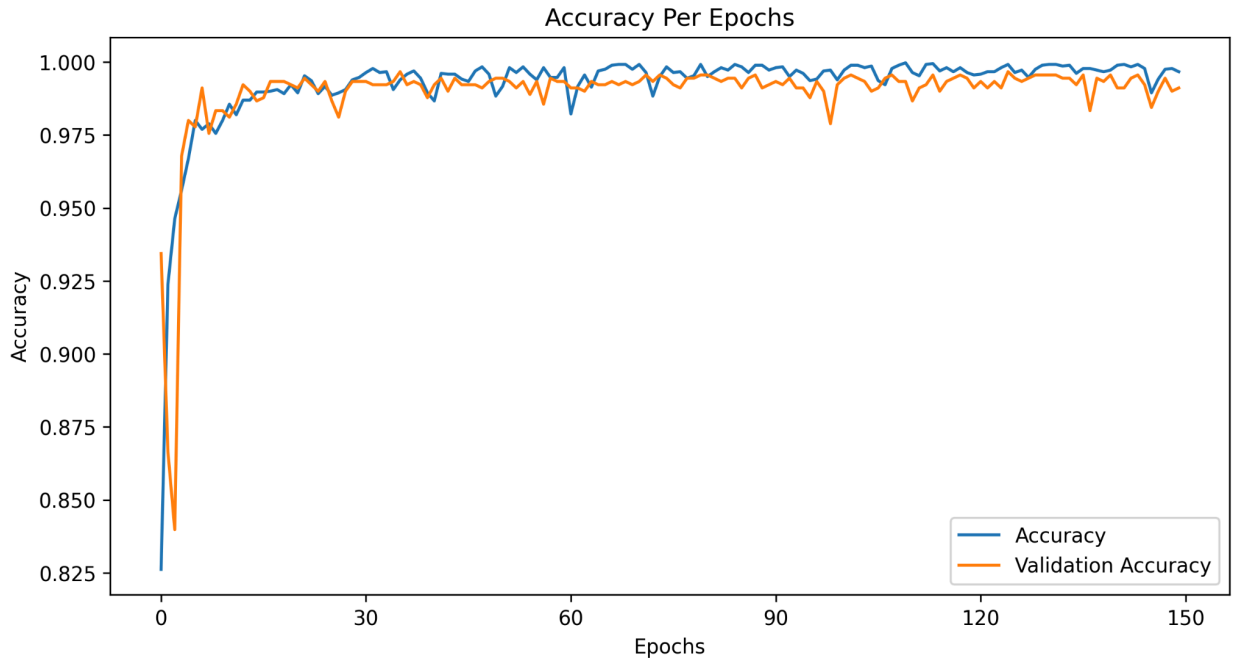

Figure 7: Accuracy Per Epochs for Conv1D Model

6 Discussion

6.1 Answering the Research Question

The results strongly support the hypothesis that a machine learning model, which is Conv1D, can effectively classify ASMR-style audio from common audio. The high accuracy and low loss values achieved across both training and validation datasets suggest that the model has successfully learned to distinguish the subtle acoustic features characteristic of ASMR-style audio.

The Conv1D model's success shows the potential of using machine learning methods for ASMR audio classification. It indicates that ASMR-style audio has unique sound characteristics that machine learning models can detect and leverage.

6.2 Limitations

Despite the promising results, there are several limitations to this study. One limitation is the potential overfitting observed in the later epochs, as evidenced by the minor fluctuations in validation loss and accuracy. This suggests that while the model performs well on training data, its performance on unseen data may not be as accurate. This issue may be attributed to the limited computational resources and time available for training, resulting in a training dataset that may not be large enough to fully capture the variability in ASMR-style audio. The dataset used for training and validation may not encompass the full diversity of ASMR-style sounds, which could affect the model's generalizability to other types of ASMR-style audio not represented in the dataset. Observations indicate that ASMR content creators are predominantly female, which is reflected in the dataset used. This gender imbalance could introduce bias into the classification model, potentially impacting its effectiveness in classifying ASMR-style audio produced by male creators.

Another limitation is the use of a single type of neural network architecture (Conv1D). While Conv1D proved effective, other architectures might capture different aspects of the audio signals that were not fully exploited in this study. For example, a Conv2D architecture could take advantage of the spatial relationships in the audio data, while a recurrent neural network (RNN) could effectively model the sequential dependencies inherent in the audio samples, shedding light on temporal patterns that may not be fully used by the Conv1D model.

6.3 Future Research

The minor fluctuations in validation loss and accuracy, particularly in the later epochs, indicate areas for potential improvement. Future research could concentrate on integrating regularization techniques like dropout or L2 regularization to further reduce the slight overfitting observed and enhance the model's resilience. Exploring additional sound characteristics or combining several feature extraction methods might enhance the model's capability to grasp the subtleties of ASMR-style audio. Experimenting with more intricate neural network architectures such as recurrent neural networks (RNNs) or transformer-based models may potentially result in even higher accuracy and improved generalization abilities.

The development of this classification model can contribute significantly to future research on ASMR-style audio synthesis. The study's findings suggest that synthesizing ASMR-style audio could become viable in future, enabling individuals to more easily select and customize their preferred ASMR content. By automating the classification of ASMR-style audio, this research not only enhances content curation on streaming platforms but also contributes to the broader field of audio classification and voice technology. This automation could lead to the creation of personalized ASMR experiences, tailored to individual preferences, and enhance the user experience on various digital platforms.

Also, future studies could explore the integration of multimodal approaches, combining audio with other sensory inputs like visual and tactile feedback, to create more immersive ASMR experiences. Investigating the psychological and physiological impacts of ASMR-style audio could also provide deeper insights into its effectiveness and potential therapeutic applications.

7 Conclusion

The Conv1D model developed in this study demonstrates a high level of effectiveness in classifying ASMR-style audio, supporting the proposed hypothesis and offering a promising tool for further exploration and application in the field of ASMR-style audio analysis. The high accuracy and low loss values achieved across both training and validation datasets indicate that the model successfully captures the subtle acoustic features characteristic of ASMR-style audio, validating the potential of machine learning methods for this purpose.

The implications of this study extend beyond mere classification. The development of an effective ASMR-style audio classifier can significantly enhance content curation on streaming platforms, providing users with more tailored and satisfying ASMR experiences. It also opens up new avenues for future research, including the synthesis of ASMR-style audio and the exploration of more complex neural network architectures and feature extraction methods.

Despite the study's limitations, such as potential overfitting and dataset biases, the findings lay a solid foundation for ongoing research and improvements in the model. Addressing these limitations through future work will likely lead to even more robust and generalizable models, capable of capturing a wider array of ASMR-style audio characteristics and applications.

In conclusion, this study not only achieves its primary goal of developing an effective ASMR-style audio classification model but also contributes to the broader understanding and technological advancement in the field of voice technology. The insights gained from this research pave the way for innovative applications and further scientific exploration, underscoring the importance and potential of ASMR-style audio analysis in both academic and practical contexts.

References

Arora, P., & Haeb-Umbach, R. (2017, October). A Study on Transfer Learning for Acoustic Event Detection in a Real Life Scenario. *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-6). IEEE.*

Aytar, Y., Vondrick, C., & Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems, 29.*

Dieleman, S., & Schrauwen, B. (2014, May). End-to-end learning for music audio. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6964-6968). IEEE.*

Gong, Y., Chung, Y.A., & Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778.*

Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. *2017 ieee international conference on acoustics, speech and signal processing (icassp) (pp. 131-135). IEEE.*

Imran, M.S., Rahman, A.F., Tanvir, S., Kadir, H.H., Iqbal, J., & Mostakim, M. (2021, January). An Analysis of Audio Classification Techniques using Deep Learning Architectures. *2021 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 805-812). IEEE.*

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M.D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2880-2894.*

Kumar, K. (2023). Audio classification using ML methods. *REVA University, Bengaluru, India*.

Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. *Advances in neural information processing systems, 22.*

Paulin, H., Milton, R.S., & JanakiRaman, S. (2018). Efficient Pre-Processing of Audio and Video Signal Dataset for Building an Efficient Automatic Speech Recognition System. *International Journal of Pure and Applied Mathematics, 119(16), 1903-1910.*

Pérez Zarazaga, P., Henter, G.E., & Malisz, Z. (2023). A processing framework to access large quantities of whispered speech found in ASMR. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.*

Piczak, K.J. (2015, September). Environmental sound classification with convolutional neural networks. *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP) (pp. 1-6). IEEE.*

Poerio, G.L., Blakey, E., Hostler, T.J., & Veltri, T. (2018). More than a feeling: Autonomous sensory meridian response (ASMR) is characterized by reliable changes in affect and physiology. *PloS one, 13(6), e0196645.*

Salamon, J., Jacoby, C., Bello, J.P., & Ellinger, B. (2014, November). A dataset and taxonomy for urban sound research. *Proceedings of the 22nd ACM international conference on Multimedia (pp. 1041-1044).*

Song, M., Yang, Z., Parada-Cabaleiro, E., Jing, X., Yamamoto, Y., & Schuller, B. (2023). Identifying languages in a novel dataset: ASMR-whispered speech. *Frontiers in Neuroscience, 17, 1120311.*

Sprengel, E., Jaggi, M., Kilcher, Y., & Hofmann, T. (2016). Audio based bird species identification using deep learning techniques. *LifeCLEF 2016, 547-559.*

Valente, F., & Hermansky, H. (2008). On the combination of auditory and modulation frequency channels for ASR applications.

Wu, Y., Mao, H., & Yi, Z. (2018). Audio classification using attention-augmented convolutional neural network. *Knowledge-Based Systems, 161, 90-100.*