The Iby and Aladar Fleischman Faculty of Engineering
Tel Aviv University

Accelerated Speaker Identification using Mel Spectrograms and EfficientNet

Roy Milshtein and Inbal Shalit

Roymilshtein@mail.tau.ac.il, Inbalshalit@mail.tau.ac.il

Abstract - This paper explores the potential application of EfficientNet, a convolutional neural network architecture originally devised for image classification, within the domain of speaker identification. Specifically, we investigate the adaptation of EfficientNet to process spectrogram representations of audio data obtained through Mel-Spectrogram. Our study aims to assess the efficacy of EfficientNet in accelerating speaker identification processes compared to traditional methods. Through a series of experiments, we evaluate the performance of EfficientNet in terms of accuracy, computational efficiency, and scalability. Additionally, we analyze the robustness of EfficientNet across various datasets and discuss its potential for real-world applications in audio processing tasks, such as voice-controlled systems and voice authentication. Overall, this research contributes to advancing the understanding of deep learning approaches in speaker identification and highlights the practical implications of utilizing EfficientNet in audio processing domains.

1 Introduction

Speaker identification plays a crucial role in various applications, ranging from security systems to voice-controlled devices. The ability to recognize speakers accurately and efficiently from audio data is essential for enhancing security measures, personalizing user experiences, and enabling seamless interactions with technology. Traditional methods for speaker identification often rely on manual feature extraction techniques followed by classification algorithms. However, with the advent of deep learning technologies, there has been a growing interest in leveraging convolutional neural networks (CNNs) for automated feature learning and classification tasks.

One promising neural network architecture for such tasks is EfficientNet, initially devised for image classification tasks. Its efficient design and robust performance in visual recognition have prompted researchers to explore its potential application in non-visual domains, including audio processing. In this study, we investigate the adaptation of EfficientNet to process spectrogram representations of audio data obtained through Mel-Spectrogram.

The research aims to address the limitations of traditional speaker identification methods, such as computational complexity and scalability, by evaluating the efficacy of EfficientNet in accelerating speaker identification processes. By leveraging the capabilities of EfficientNet, we seek to improve the efficiency and accuracy of speaker identification systems, thereby enhancing their practical utility in real-world applications.

EfficientNet models display varying success rates in speaker identification tasks. Despite smaller models proving competitive with larger ones, the anticipated performance improvement with larger models hasn't materialized.

As speaker identification becomes increasingly pivotal in diverse applications such as security systems and voice-controlled on edge AI devices, the exploration of neural network architectures like EfficientNet presents a promising avenue for enhancing accuracy and efficiency in this domain.

2 Related Work

The approach undertaken in this project exhibits both similarities and differences with the referenced works. Both approaches aim to exploit convolutional neural networks' capabilities to enhance speaker identification accuracy and efficiency. However, a key distinction lies in the preprocessing of audio data. While the referenced works primarily focus on processing raw audio waveforms directly, this project diverges by transforming the audio data into mel-spectrograms. These spectrogram representations serve as input to the EfficientNet, effectively treating the speaker identification task as an image classification problem. This approach leverages the strengths of both image processing and deep learning, potentially offering advantages in feature extraction and model performance. Therefore, while sharing the overarching goal of improving speaker identification systems, this project introduces a novel methodology that integrates mel-spectrograms and EfficientNet, setting it apart from the referenced works.

3 Dataset

In this project, the “LibriSpeech” dataset is utilized to enhance speaker identification using EfficientNet on Mel-Spectrograms. The “LibriSpeech” dataset, widely employed in automatic speech recognition research, comprises a substantial collection of English speech recordings sourced from audiobooks. Specifically, it encompasses approximately 1,000 hours of speech data acquired at a 16 kHz sampling rate, categorized into clean and other segments. In this study the clean set containing 360 hours of speech data was used. The set is gender-balanced with 921 speakers, and each speaker is limited to 25 minutes of audio. The dataset was split into train (70%) validation (10%) and test (20%), where the division is done per speaker, so there are recordings of each speaker in every set.

Prior to its integration into our project, preprocessing steps were undertaken to ensure data integrity and compatibility with our model architecture. This involved standardizing audio formats, applying noise reduction techniques, and segmenting recordings into manageable units for analysis.

4 Data Preprocessing

The data we examine is raw audio, but the network accepts image-like inputs, therefore the main challenge is to find the best way to convert our signal to a fitting input.

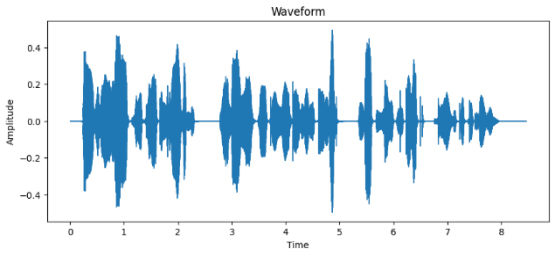

Figure 1: Waveform of a raw audio

To address the challenges outlined in the introduction regarding speaker identification, several approaches were considered and evaluated.

4.1 STFT

Initially, a visual representation of the short-time Fourier transform (STFT) was attempted as a means of preprocessing the audio data. The STFT is another variant of the Fourier Transform that breaks up the audio signal into smaller sections by using a sliding time window. It takes the FFT on each section and then combines them. It is thus able to capture the variations of the frequency with time.

$$(1)\ STFT\{x[n]\}(m,\omega) = X(m,\omega) = \sum_{-\infty}^{\infty} x[n] \cdot \omega[n-m] e^{-j\omega n}$$

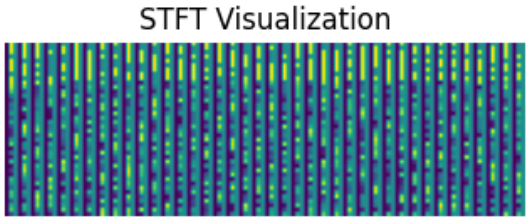

Figure 2: STFT of an audio sample

However, this approach yielded unsatisfactory results, as it failed to adequately capture the relevant features for speaker identification.

4.2 Mel – Spectrograms

Subsequently, the focus shifted towards utilizing mel-spectrograms, a commonly employed technique in audio signal processing. Mel-spectrograms offer a visual representation of the frequency content of audio signals [1], with emphasis on perceptually relevant features, by using a non-linear scale for the frequencies. This method also includes conversion to a logarithmic scale for the amplitude of the frequencies, therefore the scale is in decibels. This approach showed promise in improving the performance of speaker identification tasks compared to the initial STFT representation.

$$(2)\ M(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

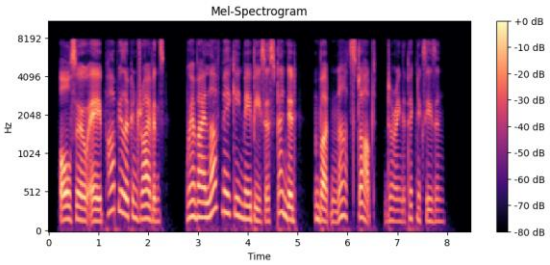

Figure 3: Mel – Spectrogram Corresponding to the Raw Audio

Further experimentation was done through Mel-frequency cepstral coefficients (MFCCs), that focus on human speech frequencies.

$$(3)\ MFCC_i = \sum_{k=1}^{N} \log(Mel_k) \cdot \cos\left(\frac{\pi(k-0.5)}{N}\right)$$

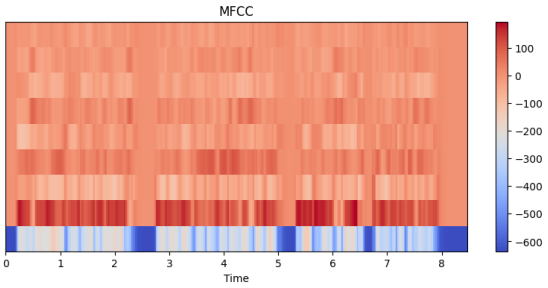

Figure 4: MFCC Corresponding to the Raw Audio

This method takes the MelSpectogram coefficients and applies further processing steps to better fit human speech [2]. This selects a compressed representation of the frequency bands from the MelSpectogram that correspond to the most common frequencies at which humans speak.

The MFCC extracts a much smaller set of features from the audio that are the most relevant in capturing the essential quality of the sound. This appeared to be problematic in training due to the big loss of data from the sample and resulted in poor performance.

Therefore, the method selected is the MelSpectogram, that encapsulates the most amount of relevant data compared to the other methods.

4.3 Data augmentation

A common technique to increase the diversity of the dataset, particularly when there is not enough data, is to augment your data artificially. We do this by modifying the existing data samples in small ways. For instance, with images, we might do things like rotate the image slightly, crop or scale it, modify colors or lighting, or add some noise to the image.

Just like with images, there are several techniques to augment audio data as well. This augmentation can be done both on the raw audio before producing the spectrogram, or on the generated spectrogram.

Attempts of spectrogram augmentation were made during the experiments that included frequency masking and time masking, and both

delivered poor results. Future attempts could be made using raw audio augmentation techniques such as pitch shift, time stretch, or adding noise.

5 Proposed Method

EfficientNet, a convolutional neural network architecture originally devised for image classification, was selected as the primary model for speaker identification tasks. Originally released by Google AI [3], EfficientNet is designed for high accuracy with fewer parameters and computational resources. By balancing the depth, width, and resolution of the network using a compound scaling method, EfficientNet achieves efficiency without sacrificing performance. It has since become widely adopted across various tasks like image classification, object detection, and semantic segmentation.

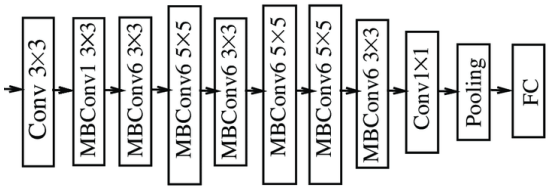

Figure 5: EfficientNet b0 architecture [3]

The proposed methodology for the experiment is based on the scalability of EfficientNet from B0 up to B7, growing the network complexity to improve accuracy.

CNN Network	Number of Parameters
EfficientNet-B0	5.3M
EfficientNet-B1	7.8M
SincNet	9.17M
EfficientNet-B2	9.2M
EfficientNet-B3	12M
SincSquareNet	12.6M
EfficientNet-B4	19M
SincNet and SincsqaureNet fusion with Attention	21.9M
EfficientNet-B5	30M
EfficientNet-B6	43M
RANet	45.3M
EfficientNet-B7	66M

Table 1: parameter count for different nets [4][5]

When comparing to the nets shown in the referenced work, the smaller EfficientNets have significantly less trainable parameters. The aim is to achieve a similar or better accuracy rate with simpler nets.

The decision to prioritize EfficientNet alongside MelSpectogram was based on empirical evaluation, which demonstrated the potential of both approaches in improving speaker identification performance. Additionally, the integration of EfficientNet aligns with recent advancements in deep learning architectures and offers potential benefits in terms of computational efficiency and model scalability.

6 Experiments

This section delves into the practical aspects of our experiments, focusing on input data preparation, challenges in data augmentation, training the networks, the significance of fine-tuning parameters, and the results.

The experiments are conducted on a CONDA python 3.10 platform running on a PC equipped with an AMD Ryzen 7 5800X processor, 32 GB of RAM, and an NVIDIA RTX 3060 graphics card (12 GB GPU).

6.1 Training process

The training process involved starting with the EfficientNet-B0 architecture and progressively scaling up the model to achieve higher accuracy. However, scaling up the model introduced challenges related to computational resources and training time. Fine-tuning the

model parameters and optimizing the training process for larger models proved to be more complex than expected.

Larger models encountered more prominent overfitting and required stronger regularization for training, as described below.

6.2 Network Input

Converting audio to image with the correct dimensions required many calculations and lots of trial and error. When calculating the MelSpectogram for a given input, the parameters affect the shape of the output image. After finding the fitting parameters there is still a problem with the dimension due to the RGB nature of images. The best result was stacking the spectrogram 3 times to fit the input layer.

Audio sample duration was an important parameter to consider when training and showed to affect the result significantly. The optimal duration used for the training was 3 seconds. Longer samples were truncated, and shorter samples were zero padded.

Further shaping was required to accommodate different lengths of audio samples. For shorter samples, zero padding was used before the transformation, and for longer samples the spectrogram was truncated to the specified length.

6.3 Hyperparameters

Due to the unique approach of audio to image conversion for the network, the tuning of hyperparameters proved to be a major obstacle in training. A very small learning rate was key to convergence, whereas too large would lead to a fast divergence between train and validation accuracy. The final learning rate selected is 0.00002.

The batch size not only affected the convergence rate, but also the speed of training on the GPU. When selecting a batch size too big (over 100), the data did not fit into memory and the training was orders of magnitude slower due to the swapping between memory and disk. The selected batch size of 80 was chosen as the maximum size to fit the GPU memory without hurting performance.

Dropout played a crucial role in mitigating overfitting during the experiments. Initially set to a low value of 0.05 for the EfficientNet-B0 architecture, dropout effectively prevented the model from overfitting to the training data, and still managed to converge. However, as we scaled up to larger EfficientNet models, we encountered challenges in determining the optimal dropout rate. Increasing the dropout rate to counter overfitting in larger models proved to be a delicate balance, as excessively high dropout rates hindered the model's ability to learn discriminative features effectively. Despite efforts to find the optimal dropout rate for larger nets, the results were not conclusive, and more testing should be done to find the optimal value for larger nets.

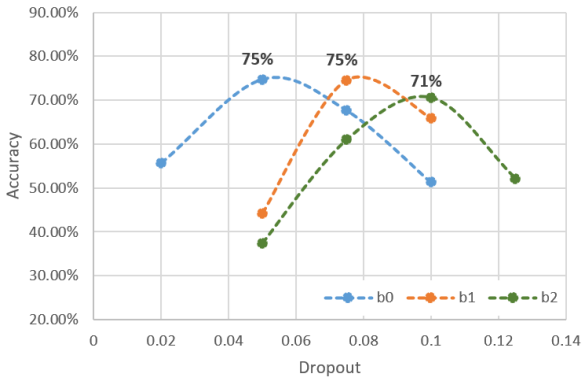

Figure 6: Dropout vs. Accuracy during training for different nets

Data augmentation for audio presented unique challenges compared to images. While techniques such as rotating, cropping, or adding noise to images are well-established, using analogous operations for audio signals required careful consideration. All testing with any amount of data augmentation proved to be counterproductive, and severely hurt the results. The dataset contains 360 hours of audio split over 104,014 samples therefore, data augmentation was not necessary to enlarge a limited dataset.
Further analysis should be done on applying augmentation directly to the raw audio versus the generated spectrogram.

Training utilizes an Adam optimizer, and weight decay of 0.001 in all training.

6.4 Results

Training EfficientNet-B0 up to B2 with the best parameters specified above, yielded the following results:

EfficientNet	Validation Accuracy	**Test Accuracy**
B0	75.13%	**74.71%**
B1	73.19%	**74.52%**
B2	75.66%	**70.57%**

Table 2: EfficientNet models Accuracy during training

Comparison to other models from related works yielded the following graph:

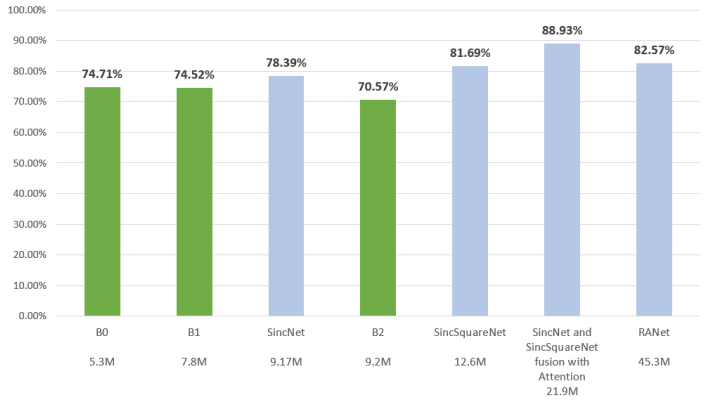

Figure 7: EfficientNet Accuracy comparison in speaker recognition [4][5]

Each column states the network test accuracy results. The nets are sorted by parameter count, as stated below each net.

7 Discussion

Our experimentations yielded varying levels of performance in speaker identification tasks. Notably, both the EfficientNet B0 and B1 models achieved a commendable test accuracy of 75%. However, the performance of the B2 model exhibited a slight decline, achieving a test accuracy of 71%. This unexpected drop in performance for the B2 model warrants further investigation to identify potential underlying factors.

The decline in performance observed with the EfficientNet B2 model could potentially be attributed to its increased complexity relative to the dataset size and task requirements. The higher capacity of the b2 model may have led to overfitting or optimization challenges during training, hindering its ability to generalize effectively to unseen data. Additionally, the diminishing marginal utility of larger models suggests that the increased model capacity may not necessarily translate into proportional improvements in performance, especially if the dataset does not fully exploit the model's capabilities.

7.1 Comparison to other nets

Comparing the findings with those reported in referenced articles, particularly concerning SincNet architectures, provides valuable insights into the effectiveness and potential limitations of EfficientNet for speaker identification tasks. Despite achieving a lower accuracy with EfficientNet-B0 model (75%) compared to other models, its performance relative to parameter efficiency is noteworthy; 3% accuracy difference with 42% fewer parameters than SincNet, suggests significant potential for surpassing SincNet's performance using simpler neural network architectures.

This demonstrates the potential for further research to enhance the basic EfficientNet architecture through parameter tuning, attention mechanisms, or other enhancements to bolster its effectiveness in speaker identification tasks.

Furthermore, our results indicate that the performance did not improve with larger EfficientNet models, necessitating a deeper investigation into the underlying reasons to fully realize the scalability potential of EfficientNet for audio processing tasks. Addressing these challenges could position EfficientNet as a credible alternative to SincNet, offering improved efficiency and scalability while maintaining competitive performance levels.

7.2 Future directions

To further enhance the performance of EfficientNet in speaker identification tasks, several research avenues warrant exploration. One promising direction is to investigate optional enhancements to the EfficientNet-b0 architecture. Integrating attention mechanisms or employing self-distillation techniques could potentially improve the model's ability to extract relevant features from audio data. Additionally, fine-tuning hyperparameters and incorporating additional layers or modules may yield better results, thereby enhancing the overall performance of the model.

Optimizing the larger EfficientNet models, such as b1 and b2, is another critical area of focus. These models have the potential to surpass the performance of b0, but achieving this requires careful tuning. Experimenting with different hyperparameters and regularization techniques can help prevent overfitting and improve generalization. Addressing these optimization challenges is essential for harnessing the full capabilities of the larger models.

More future research could be done in alternative methods for converting audio signals to image representations. Techniques such as wavelet transforms or advanced spectrogram variations could capture more relevant features for speaker identification, compared to the used mel-spectrograms.

Incorporating self-attention mechanisms within the EfficientNet models is another avenue that could improve performance. Self-attention can enhance the model's ability to focus on crucial parts of the input data, thereby improving feature extraction and classification accuracy.

Leveraging transfer learning from pre-trained models on similar tasks or datasets could also provide a significant performance boost. Fine-tuning these pre-trained models on speaker identification tasks allows us to capitalize on existing knowledge, potentially improving the results without extensive training from scratch. Transfer learning has been shown to be effective in various domains, and its application to speaker identification could yield similar benefits.

Finally, investigating optimal sampling strategies for the audio data used in spectrogram generation can enhance the quality of input features. Experimenting with different sampling rates and segment lengths may help capture the most relevant audio characteristics, leading to improved model performance. Optimizing the audio sampling process ensures that the input data is of the highest quality, providing a solid foundation for accurate speaker identification.

By pursuing these research directions, the full capabilities of EfficientNet for speaker identification tasks can be realized, potentially achieving performance levels that exceed current benchmarks. Addressing these challenges and opportunities will contribute to the advancement of deep learning approaches in audio processing and speaker identification.

8 Conclusions

This study demonstrates the potential of EfficientNet architectures for speaker identification tasks, showcasing notable accuracy with reduced parameter counts compared to traditional methods. By adapting EfficientNet to process Mel-Spectrogram representations of audio data, we highlighted its efficiency in handling audio processing. The broader implication of our research lies in its contribution to advancing deep learning approaches in speaker identification, suggesting that simpler neural network architectures can achieve competitive performance. Our findings pave the way for future explorations to further optimize and enhance EfficientNet, positioning it as a viable and efficient alternative to existing models in the field of audio processing and speaker recognition.

9 Bibliography

[1] L. Roberts, "Understanding the Mel Spectrogram," *Analytics Vidhya*, Mar. 6, 2020. [Online]. Available: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[2] E. Deruty, "Intuitive understanding of MFCCs," *Medium*, Sep. 16, 2022. [Online]. Available: https://medium.com/@derutycsl/intuitive-understanding-of-mfccs-836d36a1f779

[3] Tan, M., & Le, Q. v. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. 36th International Conference on Machine Learning, ICML 2019, 2019-June.

[4] B. Saritha, R. H. Laskar, M. Choudhury, and A. Monsley K, "Optimizing Speaker Identification through SincsquareNet and SincNet Fusion with Attention Mechanism," *Procedia Computer Science*, vol. 233, pp. 215-225, 2024.

[5] B. Saritha, M. A. Laskar, R. H. Laskar, and M. Choudhury, "Raw Waveform Based Speaker Identification Using Deep Neural Networks," in *2022 IEEE Silchar Subsection Conference (SILCON)*, Silchar, India, 2022, pp. 1-4.

10 Appendix

The Git folder contains:

EfficientNet_Speaker_Identification_Train.ipynb – notebook with full training code. Shows the training on the final b2 net. The code is split into blocks and contains many comments explaining its functions.

Saved_weights – folder with the best trained weights for models: b0, b1, b2.

Audio2img.ipynb notebook – compiles methods of converting audio to an image. Contains visualizations for a sample.

LibriSpeech – dataset folder with smaller sets not used in the final project. The full dataset used is the train-clean-360 that has 30GB and is too big for the drive.

The code was trained on a PC with windows, and not on Google Colab, therefore adaptations must be made to file paths in order to run. The dataset is very large so running on Colab was not a viable option.

In order to test the saved weights, run all the blocks, except for the training block. This will define the models, load the data, load the saved weights, and finally evaluate the model on the test set.