Baby Cry Detection

[*]Note: Sub-titles are not captured in Xplore and should not be used

***Abstract*—This project aims to enhance the functionality of baby cry analysis models by improving the accuracy of detecting baby cries amid various other sounds and background noises. Similar to how voice-activated assistants like Amazon's Alexa or Apple's Siri utilize advanced algorithms to accurately detect and respond to specific voice commands amidst a multitude of background noises (e.g. 'Hey Siri'), we aim to develop a highly accurate cry detection algorithm. Cry detection is eventually used to enhance cry analysis algorithms, which distinguish between different types of baby cries—such as those resulting from hunger, pain, or discomfort—with precision. By achieving this level of accuracy, we can provide caregivers and health professionals with critical insights into the immediate needs of infants, leading to more timely and appropriate responses.**

***Index Terms*—acoustic scene detection, cry detection**

I. Introduction

Building robust models that discern audio features in domestic environments amongst a variety of different sounds can be a challenging task. A common problem in this use case is the lack of quality annotated data available for some more specific use cases, such as baby cries. This is addressed by manual annotation of reliable data, and data augmentation techniques to pragmatically increase the robustness of baby cry analysis models, as extracting the cry interval only proves useful as it negates the amount of unwanted noise that affects the models that perform the analysis as to declare the semantics behind the baby cries.

II. Related work

- Current cry detection models only classify a full audio example, rather than intervals within the example, which may affect analysis outcomes.
- Current methods use a simple CNN-BiLSTM as a less computationally expensive option, and state of the art use Audio Spectrogram Transformers, which require much more compute and data.

A. Cry detection works

III. Datasets

- The data obtained initially was unlabelled. It was then annotated using "Praat" such that every audio sample had 3 separate annotations from different annotators.
- The data was then padded to 8 seconds, with it being cut into 160 intervals, each of 0.05 seconds.
- 457 samples were annotated by the team, and 11 samples from DCASE2023 were used as extra negative samples
- The 457 baby cry samples consisted of 276 boys and 181 girls, and a large majority were less than 24 months old.

A. Our Method

- In the real-time baby cry audio segmentation task, a Spectrogram-based MobileNetV2 model achieved superior performance compared to other evaluated models. This method attained a frame accuracy of 0.92, a frame-based F1 score of 0.87, and an event-based F1 score of 0.87, demonstrating the best performance across all three metrics. Spectrogram-based MobileNetV2 converts WAV audio files into spectrograms using Mel-frequency cepstral coefficients (MFCC) for subsequent classification via a pre-trained MobileNetV2 model.
- We employed a dataset of 374 baby cry audio samples for training and 94 samples for testing. The batch size was set to 8, and the model was trained for 10 epochs. Leveraging the computational efficiency of the MobileNetV2 architecture, the training process on a GPU achieved exceptionally fast speeds, completing in under 30 seconds.
- We adopted three standard evaluation metrics to assess the performance of the models: frame accuracy, frame-based F1 score, and event-based F1 score. Frame Accuracy represents the standard accuracy calculation, reflecting the percentage of correctly classified frames within the audio samples. Given the model outputs a label for each frame, the frame-based F1 score is the harmonic mean of precision and recall calculated at the frame level. Event-based F1 Score is specifically tailored for audio segmentation tasks involving overlapping sound events, as described in the paper "Metrics for Polyphonic Sound Event Detection" by Mesaros et al. [1]. We propose an algorithmic method to calculate true positives, false positives, and false negatives based on the number of detected sound events, predicted events not present in the ground truth, and undetected events. These values are then used to compute the event-based F1 score. The rationale behind employing this metric lies in our observation that high frame accuracy for certain methods might not necessarily translate to accurate audio segmentation. Therefore, the event-based F1 score provides a more robust assessment by focusing on correctly identified sound events.

Identify applicable funding agency here. If none, delete this.

IV. Experiments & Results

- For a thorough evaluation of the Spectrogram-based MobileNetV2 model's performance, we compared it against five alternative architectures: Bidirectional LSTM, Transformer (with tabular features), Transformer (with spectrogram features), CNN-Transformer, and Tinyformer. These models were each examined under two baseline configurations differing in data processing techniques. Both configurations employed Mel-frequency cepstral coefficients (MFCC) for signal processing. However, one configuration converted each time unit of the audio into 32 MFCC coefficients, while the other converted each time unit into a spectrogram. To ensure consistent evaluation, all models were assessed using the three aforementioned metrics: frame accuracy, frame-based F1 score, and event-based F1 score.

TABLE I
Model Evaluation

Models	Frame-based F1 Score	Event-based F1 Score	Accuracy
BiLSTM	0.76	0.64	0.83
TinyFormer	**0.87**	**0.87**	0.91
Transformer (MFCC)	0.76	0.76	0.84
CNN Transformer (Spect.)	0.85	0.85	0.89
MobileNetV2	**0.87**	**0.87**	**0.92**
Transformer (Spec.)	0.82	0.82	**0.92**

- The results reveal that Tinyformer and Spectrogram-based MobileNetV2 achieved the highest performance in terms of both frame-based and event-based F1 scores. Conversely, Bidirectional LSTM (BiLSTM) and Transformer models employing tabular features exhibited the lowest performance on these metrics. Notably, Spectrogram-based MobileNetV2 maintained the top position for frame-based accuracy, while BiLSTM remained the least accurate. These observations suggest the superiority of visual features over tabular features for this task. Vision models like MobileNetV2, by leveraging spectrograms, effectively capture temporal information from the audio data, leading to improved segmentation performance.
- Notes for this section: state of the art models like MarbleNet and AST use a 16k sample rate for this model, but all of our annotated data is 8k. Resampling our data to 16k resulted in subpar results.
- Initially, experiments were planned around state-of-the-art models like the Audio Spectrogram Transformers which are available publicly, but they use 16k sampling rate, while our data was 8k. Resampling to 16k caused a lot of issues with noise, drastically reducing the performance using pretrained state of the art. Using only our data on simpler models such as the CNN

V. Discussion & Conclusion

- The Spectrogram-based MobileNetV2 model demonstrates promising capabilities for real-time baby cry audio segmentation, excelling in both theoretical and practical aspects. Theoretical Validation: The model achieves strong performance across all three evaluation metrics (frame accuracy, frame-based F1 score, and event-based F1 score), signifying its effectiveness in accurately segmenting baby cry audio data. Real-World Applicability: The model exhibits robustness in real-world scenarios by efficiently detecting baby cries amidst various noise environments. Additionally, it successfully differentiates baby laughter and adult crying sounds, indicating its ability to function at a product-level capacity.

While the Spectrogram-based MobileNetV2 model demonstrates promising results, there are areas for improvement to enhance its robustness and generalizability.
- Data Imbalance: The training data exhibits a class imbalance, with baby cry segments constituting a smaller portion compared to non-cry audio. This imbalance could potentially bias the model towards the majority class. To address this, we will explore techniques like data augmentation or cost-sensitive learning in future iterations.
- Model Complexity: The pre-trained MobileNetV2 model possesses a relatively high level of complexity for this specific task. Investigating the use of simpler models specifically designed for audio classification could potentially improve efficiency while maintaining acceptable performance.
- Data Enhancement and Loss Optimization: Further research will focus on advanced data enhancement techniques to potentially enrich the training data and improve model generalization. Additionally, exploring alternative loss functions tailored for audio segmentation tasks could potentially lead to further performance optimization.

References

[1] Mesaros A, Heittola T, Virtanen T. Metrics for Polyphonic Sound Event Detection. Applied Sciences. 2016; 6(6):162.