UNIVERSITY OF CALIFORNIA SAN DIEGO

Machine Learning-based Classification of Infant Directed Speech in Multiple Languages

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Electrical Engineering (Intelligent Systems, Robotics, and Control)

by

Israa Aljarb

Committee in charge:

Professor Edward Wang, Chair
Professor Yuanyuan Shi
Professor Xiaolong Wang

2023

The Thesis of Israa Aljarb is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

I want to share my heartfelt condolences for my beloved uncle, who played a crucial role in shaping who I am today. He was more than just a family member to me; he was my mentor, guide, and confidante. He was the one person who never doubted me and always believed in me, even when I did not believe in myself. His unwavering support and encouragement gave me the strength to pursue my dreams and achieve my goals. He inspired me to leave my small hometown to pursue higher education, and his teachings continue to inspire me to this day. It saddens me that he left this world just before I began my graduate studies, but I know his spirit lives on in me. I carry his memory with me every day, and his love and wisdom continue to guide me through life's challenges.

My dearest friends, my heart aches for you. Your time on this earth was too brief, but your impact on my life will never fade. We walked hand in hand through the college years, sharing everything from laughter to hardships. Even though you left us too soon, I keep your memory alive every day and night. You taught me courage, strength, and the true meaning of friendship. Every success and obstacle I overcome is for you, my dear friends, my fallen comrades from the life book too early.

In this dedication, I honor those I lost, including my late uncle and friends. Their memories continue to inspire and guide me through difficult challenges. May their legacies live on through me and others who knew them. I am forever grateful for the moments we shared.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I want to thank my committee chair, Edward Wang, for his priceless assistance throughout my research. His expertise, patience, and determined support have motivated me to seek excellence. I sincerely appreciate his insightful feedback and constructive criticism, which have always challenged me to improve my work.

I would also like to acknowledge“Varun Viswanath“ from DigiHealth Lab as he was my first monitor who supported me throughout the initial phases of my research journey, especially during the first data collection and model-building stages. His experienced feedback played an essential role in laying the foundation for this study.

Once more, I thank the undergraduate student “Deepansha“ Singh from the Computer Science and Engineering department, who assisted me during the first phases of this research. Her assistance has been crucial in shaping the success of this project.

I genuinely appreciate all the DigiHealth lab members' collaboration and support throughout this research. Their handout and teamwork have been influential in making this effort a success.

I like to express my gratefulness to my family and friends for their constant support and understanding during this journey. Their encouragement and belief in me have been a constant source of motivation, and I am grateful for their presence in my life.

In the 2.1.1 section, I would like to thank Professor Lama K. Farran for her collaboration in the Arabic dataset, as 23.9% of the used Arabic dataset was shared by her. Also, I want to thank Elise A. Piazza for sharing with us his English dataset, where 9% of the used English dataset in this study was shared by him.

Chapters 2 and 3, will be submitted for publication of the material as it may appear in JMIR 2024 Papers. The thesis author was the primary investigator and author of this paper.

ABSTRACT OF THE THESIS

Machine Learning-based Classification of Infant Directed Speech in Multiple Languages

by

Israa Aljarb

Master of Science in Electrical Engineering (Intelligent Systems, Robotics, and Control)

University of California San Diego, 2023

Professor Edward Wang, Chair

Recent studies show that some toddlers with Autism Spectrum Disorder (ASD) do not respond to high-pitched and exaggerated intonation speech, known as infant-directed speech (IDS). Understanding the caregiver-child speech interaction system is critical to detect early signs of ASD. Therefore, this study evaluates the potential for a signal processing pipeline with deep learning in classifying IDS and adult-directed speech (ADS) in multiple languages: English, Arabic, Spanish, and Chinese. Our pipeline classifies IDS and ADS in single-language and multi-language models using 3260 ADS and IDS audio files. Results demonstrate that a classification model for single-language achieved accuracy between 85% and 94% for all

languages. The multi-language model achieved an accuracy of 93% across all language datasets. Results from this technique indicate the possibility of accurately classifying ADS and IDS. This study offers opportunities to develop caregiver-child interaction systems and increase tools for early detection of ASD in toddlers.

Chapter 1

Introduction

In this chapter, we will thoroughly examine the fundamental components that lay the groundwork for our research. Our initial step is to provide a detailed background that establishes the context for our study. Following the background, we will proceed to explain the reasons that have inspired our interest in this specific research area. Our objective is to communicate the underlying factors that have influenced the scope and importance of our study. Subsequently, we define the research problem at the heart of our inquiry. Simultaneously, we outline the specific objectives that guide our research efforts, delineating the outcomes we aspire to achieve and the questions we seek to answer. To contextualize our research within the broader academic discourse, we conduct a thorough literature review. This involves a critical examination of existing scholarly works, theories, and empirical studies related to our research area. By analyzing prior research, we position our study within the larger body of knowledge, identifying gaps that our research aims to address. Overall, this chapter serves as a comprehensive introduction.

1.1 Background

Autism spectrum disorder (ASD) is a neurodevelopmental disorder that commonly displays in early childhood (Centers for Disease Control and Prevention, 2022) [5]. According to (the American Psychiatric Association 2013) website, autism spectrum disorders contain a diverse and complex range of neurodevelopmental conditions marked by challenges in com-

munication, interpersonal relationships, and the presence of limited and repetitious behaviors and attractions[3]. Because of the significant brain development and neural plasticity during the early years of life, earlier identification of ADS is necessary (Centers for Disease Control and Prevention, 2022) [5]. According to the study instructed by (Karen et al. 2023), toddlers who were diagnosed with ASD showed notably lower attention for motherese speech compared to toddlers without ASD [22]. Again, the same analysis [22] demonstrated that toddlers who concentrated on motherese speech by or below 30% had a 94% chance of correctly having ASD; this suggests a significant connection to difficulties in socializing and communicating with others. Further, the study's findings, which were conducted by(Yaqiong et al. (2022)), early signs of autism in infants can be a lowered response to affective speech and reduced caregiver-child interactions[24]. Researchers (Marisa et al. (2017))[8] have been working on finding early identifying signs of Autism Spectrum Disorder (ASD) to improve long-term outcomes. They believe enhanced auditory processing, particularly the preference for infant-directed speech or motherese, caused by social interest, can be an early indicator of ASD risk.

Infant-directed speech, also known as Motherese or baby talk, is a way of speaking that adults, typically caregivers use to communicate with infants and young children (Weiyi Ma et al., 2011)[16]. IDS encompasses features such as a pitch, exaggerated intonation, a slower pace, and simplified language. Caregivers naturally employ Motherese across cultures as it helps capture the baby's attention and aids in early language development [16]. A study [14] (Patricia K. Kuhl et al., 1997) showed that in the early months of infants' lives, they learn their native language by listening to adults speaking sounds. This type of speech is used across different cultures and languages as the same paper [14] found that when mothers from the United States, Russia, and Sweden talk to their babies, they use more extreme vowel sounds compared to when they spoke to adults.

The above initial discoveries highlight the significance of observing whether caregivers use infant-directed speech when interacting with their child and if the child is positively responding to such language interactions. Therefore, a mother-child speech interaction system is

necessary to detect early signs of ASD. The current solution is the Language Environment Analysis System (LENA), a voice monitoring solution for babyhood to school-age children. However, a clinical provider analyzes the data offline, and LENA only provides metrics of conversational turn counts within 5-second intervals. It also does not provide information about infant-directed speech or distinguish between words and sounds at a detailed level. An ideal solution would be to offer an at-home mobile health screening for ASD. This solution can be done by using digital biomarkers to monitor mother-child interactions during infant-directed speech play, which are linked to early signs of ASD. In order to create this solution, the initial step is to develop a system that is able to identify infant-directed speech, regardless of the language being spoken.

To sum up, the importance of monitoring the interactions between caregivers and children, especially during infant-directed speech, highlights the need for more advanced screening methods. The first step is to create a system that can identify infant-directed speech, regardless of the language being spoken which can increase demand for comprehensive and easily accessible ASD screening instruments, which can help us improve early identification and intervention techniques.

1.2 Motivation

The motivation behind this research comes from the critical need to address the challenges associated with Autism Spectrum Disorder (ASD) identification, particularly in its early stages. Early detection is crucial for individuals with ASD, as it can greatly improve their long-term outcomes. Timely intervention during the formative years is especially important.

Current research findings, including studies by Karen et al. (2023) [22] and Yaqiong et al. (2022)[24], highlight the importance of observing responses to infant-directed speech (IDS) for early ASD identification. This observation, coupled with the limitations of existing monitoring solutions such as the Language Environment Analysis System (LENA), emphasizes the necessity for a more sophisticated and inclusive approach to ASD screening.

Infant-directed speech, commonly known as Motherese, serves as a valuable tool in caregiver-infant communication, playing a crucial role in language development. However, existing systems like LENA fall short of providing detailed insights into IDS, which can be early ASD identification, prompting the need for an innovative solution.

The main goal of this research is to create a machine-learning algorithm that can effectively distinguish between infant-directed speech (IDS) and adult-directed speech (ADS) in various languages. By doing so, this algorithm will not only overcome the current limitations of screening methods but also contribute to the development of a system that supports mother-child interactions during IDS play. Ultimately, this can lead to a more comprehensive, accessible, and linguistically inclusive ASD screening approach, which could transform the way we diagnose and support individuals with ASD.

Our goal is to contribute to the improvement of at-home mobile health screening for ASD by achieving these objectives. We aim to leverage digital biomarkers to enhance early detection capabilities and make a significant impact on the lives of individuals and families affected by ASD. Through this research, we are striving to promote a more proactive and inclusive approach to diagnosis and intervention. Ultimately, we are motivated to create innovative solutions that can positively impact the identification and support of ASD.

1.3 Research problem and objectives

This study aims to develop and test a machine learning algorithm that can accurately classify infant-directed and adult-directed speeches in English, Arabic, Spanish, and Chinese languages with two main objectives:

- Our initial goal is to develop a single-language model for each of the following languages: English, Arabic, Spanish, and Chinese. These models will be designed to differentiate between IDS and ADS for each language individually.

- The second goal is to develop a multi-language model that can detect ADS and IDS for all

four languages: English, Arabic, Spanish, and Chinese. This model incorporates linguistic diversity to improve speech detection and classification across various languages and dialects.

It is important to highlight that this project is the first block to create a system that facilitates mother-child interactions during infant-directed speech play, which can help detect early signs of autism spectrum disorder as the big goal of mother-child interactions is to identify early signs of ASD by observing whether caregivers use infant-directed speech and if the child positively responds to such interactions.

1.4 Literature Review

In this section, we examine the latest developments in research on machine learning techniques for analyzing infant-directed speech in real-life interactions. We provide a summary of essential papers that classify IDS and ADS based on data collection, the language used, the method employed, and the results obtained.

Takao Inoue et al. (2011)[12] examined Japanese infant-directed speech (IDS) and gathered information from 24 mothers conversing with infants between 8.1 and 9.5 months old at Nagasaki University Graduate School of Biomedical Sciences. They extracted features like Mel-frequency cepstral coefficients (MFCC), frequency (F0), and energy and used Hidden Markov Models (HMMs) to detect infant-directed speech, achieving an 84.34% accuracy rate. Second, Erika Parlato-Oliveira et al. (2019) [20] studied infant-directed speech in multiple languages: English, French, Hebrew, Italian, and Brazilian-Portuguese languages. The data was collected from different sources, including audio and video recordings. Open SMILE Low-Level features were extracted, and Support Vector Machines (SVMs) were used for classification. The accuracy of the results varied, with the lowest being 63% for Italian and the highest being 91% for Hebrew[20].

N. D. Al Futaisi et al. (2023)[1] Researchers analyzed the way English and Argentine-

Spanish speakers talk to babies and adults using different sets of data and a feature extraction toolkit. They used various techniques to classify the data and presented their results in F-1 scores ranging around 66%. Mahdhaoui and Chetouani[12] studied how infants respond to parent-infant interactions using supervised and semi-supervised learning methods. They analyzed recordings of Italian mothers interacting with their infants, identifying infant-directed speech and adult-directed speech. The researchers used various temporal, frequency, and perceptive features to extract data and used four classifiers to categorize infant-directed and adult-directed speech. Gaussian mixture models (GMM) with cepstral MFCC features were found to be the most efficient classifier.

The study[15] aimed to classify infant and parent emotional vocalizations using a dataset that included recordings from home and laboratory settings. Infant vocalizations were categorized into five types, while parent vocalizations were labeled as six types. The study evaluated three models, i.e., Linear Discriminant Analysis (LDA), two-layer Fully Connected Network (FCN), and Convolutional Neural Network with Self-Attention (CNSA). Remarkably, the CNSA model outperformed the other models and showed the best Unweighted Average Recall (UAR) on the CRIED dataset, indicating its superior performance over previous studies. The study also employed feature augmentation strategies and post-hoc feature analysis and identified acoustic features such as MFCC, log Mel Frequency Band Energy, LSP frequency, and F0, which helped in classifying vocalization types. The neural network models, especially CNSA, demonstrated promising results with an accuracy of 86.25%, Macro F1 of 68.77%, and UAR of 67.09%.

As we look at the research done before, it is evident that the main objective was to create machine-learning models to detect infant-directed speech in a single language. However, in this study, we take a significant jump forward by developing a multi-language model that is trained and tested on various languages, including English, Arabic, Spanish, and Chinese. Our comprehensive approach is crucial in leveraging the full potential of linguistic diversity. This jump enhances the model's capacity to recognize infant-directed speech across various cultural and language contexts.

Chapter 2

Methodology

In this chapter, we analyze our method of developing a machine learning algorithm that can distinguish between Infant-Directed Speech (IDS) and Adult-Directed Speech (ADS) in multiple languages, including English, Arabic, Spanish, and Chinese. We cover dataset build in terms of collecting and preprocessing. Further, signal processing and feature extraction are discussed. Furthermore, we discuss model development, including single-language and multi-language models. Finally, we will conclude this chapter by discussing the evaluation Process.

2.1 Dataset Building

One of the essential aspects of this study is creating four audio datasets for the above languages mentioned, which are collected from different places and sources. These datasets consist of both Infant-directed and adult-directed speech audio recordings.

2.1.1 Data Collection

English Datast

The English dataset was collected from three different sources, and the distribution based on each source is displayed in Figure 2.1.

English dataset collected sources are :

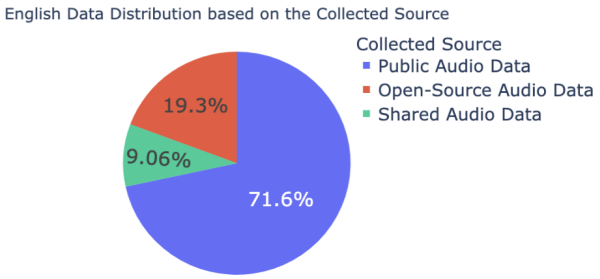


Figure 2.1. English Dataset Distribution Based on the Collected Source

Description: The pie chart illustrates the distribution of English datasets based on the collected source. The chart is divided into three segments representing different sources. Approximately 71.6% of the dataset is sourced from public datasets, 19.3% from open-source audio data, and 9% from shared audio datasets. Each segment is visually distinguished by a unique color, providing a clear representation of the dataset's source composition.

- **1:** Open source and free content websites like YouTube, TikTok, and Instagram are the main data collection sources. Our goal in using these websites was to find genuine and natural interactions between caregivers and infants. We searched for audio clips featuring a female speaking to a baby using infant-directed speech, as well as using her natural voice or adult-directed speech. We made sure to select audio clips where at least two adults were speaking to each other and a female was speaking to a baby in the same audio. We used keywords like **"women playing with baby"** and **"infant playing with mom"** to find suitable audio clips.

- **2:** The second source for the English dataset is the shared dataset from (Elise A. Piazza et al.2017))[21]. The dataset is not published, but it was shared with us upon asking. According to researchers, this dataset contains a record of 24 mothers' realistic speech while they engage with their newborns and with grown-up experimenters in their native tongue. Mothers' speeches were recorded using an iPhone and microphone in two situations: ADS and IDS. In ADS, mothers were questioned about their children's schedules, habits, and interactions. In IDS, mothers were allowed to interact with their infants naturally, playing with toys and reading board books[21].

- **3:** The final source is part of the public dataset (Courtney B. Hilton et al. 2020) [11] According to the paper, the data set comprises 1,615 sound recordings of infant-directed and adult-directed songs and speeches. 410 people from 21 different communities, ranging from urban to rural to small-scale societies, contributed to its production. The dataset has 18 languages from 11 language families. In this English dataset building, we used only infant-directed and adult-directed speeches for English speakers collected from the USA, Canada, and New Zealand.

The dataset for the English language includes three distinct accents that were gathered from three different locations. The distribution of the English language accents dataset is illustrated in Figure 2.2.

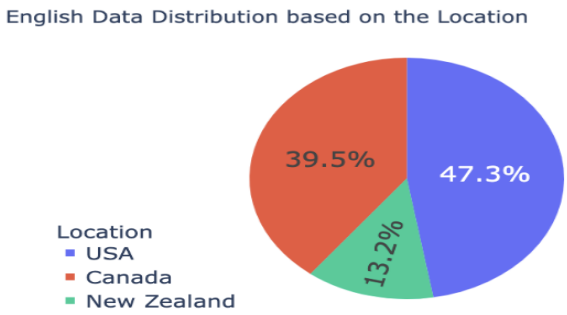


Figure 2.2. English Dataset Distribution Based on the Location

Description: The pie chart depicts the distribution of the English dataset based on its collection from three different locations. The dataset exhibits three distinct accents, with the majority collected from the USA (47.3%), followed by Canada (39.5%), and New Zealand (13.2%). Each segment is represented by a unique color, providing a visual representation of the dataset's distribution across different geographical locations.

Arabic Dataset

We obtained the Arabic dataset from two different sources, as indicated by the data distribution shown in Figure 2.3. The first source is the same as the one used for the English dataset, where we followed a similar strategy of using open-source and free content websites. For the first source, we restricted our search to locate audio files exclusively from Saudi Arabia.

The second source is the shared dataset from (Lama K. Farran et al. 2016)[7] upon asking. According to the author, this shared dataset consists of recordings of interactions between parents and infants in Lebanese-Arabic families. The recordings were taken in Lebanon, where Arabic is the primary language for most people. The records were for 12 male and 7 female infants aged $0-24$ months and their Arabic-speaking Lebanese mothers from two private and two public pediatric clinics in Lebanon. High-fidelity equipment with built-in stereo microphones was utilized to record the samples in the infants' homes. The mothers were requested to interact with their infants for 10 minutes in their usual manner at home prior to the recording sessions [7]. The Arabic language dataset contains two unlike dialects collected from two different locations. The distribution of Arabic language dataset dialects is shown in Figure 2.4

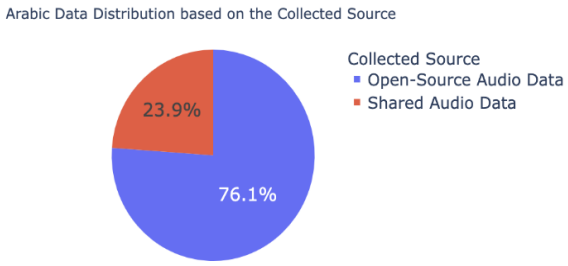


Figure 2.3. Arabic Dataset Distribution Based on the Collected Source

Description: The figure illustrates the distribution of the Arabic dataset based on the collected source. A significant portion, 76%, is sourced from open-source audio data, highlighting the reliance on publicly available resources. Additionally, 23.9% of the dataset is obtained from the shared dataset. The percentages are represented by distinct segments in the chart, providing a visual representation of the composition of the Arabic dataset with respect to different sources.

Spanish Dataset

To gather the Spanish language dataset, we utilized the first and third sources mentioned in the data collection of the English language. Specifically, in the third source [11], we only selected infant-directed and adult-directed speeches that were spoken by Spanish speakers from Afro-Colombian and Colombian Mestizo backgrounds. Figure 2.5 displays the distribution of

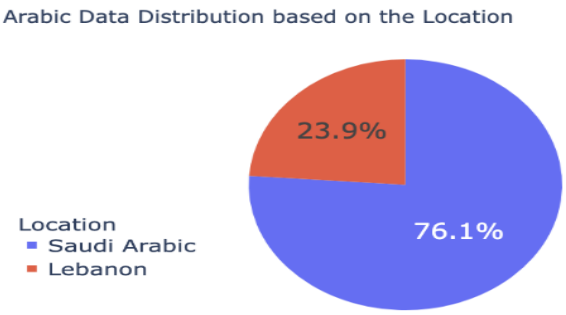


Figure 2.4. Arabic Dataset Distribution Based on the Location

Description: The figure showcases the distribution of the Arabic dataset based on the collection locations, representing two main dialects. A substantial 76% of the dataset is attributed to Saudi Arabic, reflecting a predominant source of the data. Furthermore, 23.9%of the dataset is derived from Lebanon, contributing to the diversity of dialects within the dataset. The distinctive colors in the chart visually depict the proportionate representation of each location source in the Arabic dataset.

Spanish data sources, and Figure 2.6 shows data location in the Spanish language dataset.

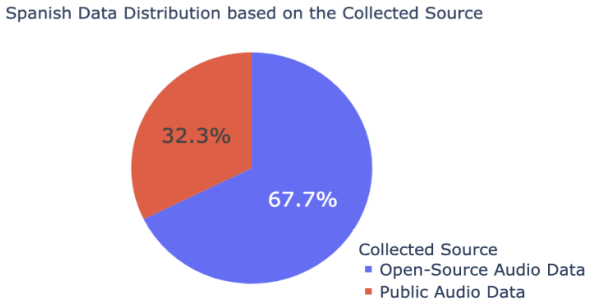


Figure 2.5. Spanish Dataset Distribution Based on the Collected Source

Description: The figure illustrates the distribution of the Spanish dataset based on the collected source. A significant majority, 67.7%, is sourced from open-source audio data, showcasing a reliance on publicly available resources for building the dataset. Additionally, 32.3% of the dataset is obtained from the public dataset. The distinctive segments in the chart visually represent the proportional contribution of each source to the overall composition of the Spanish dataset.

Chinese Dataset

We collected the Chinese language dataset from the first and third sources mentioned in the English language data collection. For the third source [11], we specifically selected infant-directed and adult-directed speeches from Chinese speakers in Beijing. In Figure 2.7, the

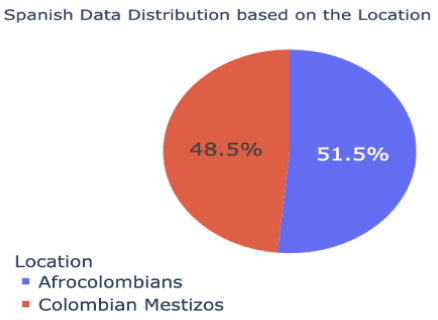


Figure 2.6. Spanish Dataset Distribution Based on the Location

Description: The figure depicts the distribution of the Spanish dataset based on the collection locations, highlighting two distinct categories. The dataset is primarily composed of 51% from Afrocolombians, reflecting a significant contribution from this particular group. Additionally, 48.5% of the dataset is derived from Colombians, showcasing the diversity of sources within the Spanish dataset. The visual representation through distinctive colors in the chart provides insight into the proportional representation of each location in the overall dataset.

data source distribution is illustrated.

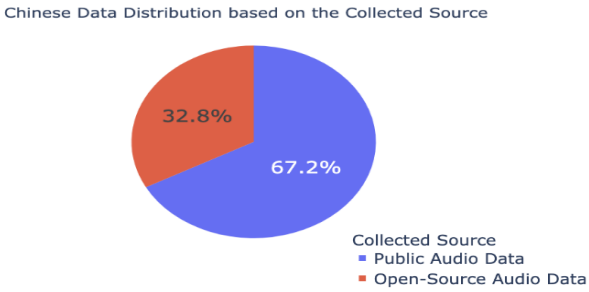


Figure 2.7. Chinese Dataset Distribution Based on the Collected Source

Description: The figure illustrates the distribution of the Chinese dataset based on the collected source. A predominant 67% of the dataset is sourced from public datasets, highlighting a substantial reliance on publicly available resources. Additionally, 32.8% of the dataset is obtained from open-source audio data. The distinctive segments in the chart visually represent the proportional contribution of each source, providing a clear overview of the composition of the Chinese dataset.

2.1.2 Data Labeling and Segmenting:

In the Data Labeling and Segmenting phase, a meticulous approach was taken to process the collected audio datasets for further analysis. After obtaining **MP3** audio files from diverse sources, Audacity, a free and versatile audio editing software, played a pivotal role in the

preprocessing stages. Ensuring consistency, each audio file was meticulously examined to guarantee the presence of a woman speaking to her baby using infant-directed speech and engaging with another adult using adult-directed speech. With a focus on six distinct sounds within each file—women speaking using infant-directed speech, women using adult-directed speech, other adult talk, baby talk, noise, and silent periods—the labeling process selectively focused on identifying and marking women speaking in the targeted speech modes. Subsequently, only the labeled segments of women speaking in the infant-directed and adult-directed speech were retained, and other sounds were disregarded. The annotating process involved a single annotator, and in cases of uncertainty between IDS or ADS, a second listener was consulted. If ambiguity persisted, the segment was subjected to a secondary evaluation, and only those segments that exhibited clarity after the second listener's judgment were retained. These labeled segments were then segmented into new audio files, saved in the **WAV** format, and organized with filenames containing the speaker ID, source information, and speaker location. Notably, audio files from specific sources, such as the third source[11] and the second source[21], had already been labeled and segmented, streamlining this particular phase. This meticulous Data Labeling and Segmenting process aimed at refining and categorizing the dataset, ensuring that only the relevant segments were included for subsequent analysis, thereby enhancing the precision and usability of the multilingual speech dataset.

2.1.3 Data Cleaning:

Following the meticulous Data Labeling and Segmenting process, a crucial step in dataset refinement involved thorough Data Cleaning. Post the segmentation of new audio WAV files, a meticulous review was conducted by re-listening to each file using Audacity. This data-cleaning phase was applied consistently across all datasets, encompassing various sources and languages. The process focused on enhancing data quality by systematically eliminating silent portions present in audio files, a process extended to intervals between words. This measure aimed to refine the dataset by ensuring that the retained audio segments were substantive and

contributed meaningfully to the dataset. Additionally, any instances of loud noise were addressed through a carefully judged reduction by two evaluators. This collaborative evaluation ensured that adjustments were made to maintain the accuracy, clarity, and cleanliness of the data. The application of Audacity in the data cleaning process provided a versatile and effective tool for refining the auditory dataset across diverse languages and sources.

2.1.4 Data Quality Assurance:

Following the completion of the labeling, segmenting, and data cleaning processes for all four language datasets, a critical quality control measure, manual verification, was implemented. To ensure the precision and accuracy of the labeled data, a native speaker proficient in each language served as an evaluator. Each evaluator meticulously listened to a subset of the corresponding language dataset, carefully judging the data quality and assessing the accuracy of the labeling. This manual verification process served as a final checkpoint, allowing for the identification and rectification of any potential discrepancies or errors that might have occurred during the labeling and cleaning phases. The involvement of native speakers in the manual verification ensured a nuanced understanding of language-specific nuances and pronunciation, contributing to the overall reliability and linguistic authenticity of the labeled datasets.

2.1.5 Data Splitting:

Upon the completion of constructing four distinct audio datasets, each comprising English (1814 seconds), Arabic (740 seconds), Spanish (992 seconds), and Chinese (1050 seconds) speech segments, our next objective was to partition each dataset into training and testing subsets. The process was executed with meticulous consideration for maintaining data independence and diversity. Employing a robust strategy, we leveraged the speaker ID, source, and location information embedded in each file's naming. Specifically, we carefully and manually separated the training and testing datasets based on the unique speaker ID, ensuring that no individual speaker's recordings appeared in both sets. This approach guarantees that each dataset possesses

Table 2.1. Sizes of Training and Testing Datasets for Each Language

Language	Number of Seconds in Training Dataset	Number of Seconds in Testing Dataset
English	1442	372
Arabic	574	166
Spanish	712	280
Chinese	840	210

The table presents the size, in seconds, of the training and testing datasets for each language in the study. The numbers represent the cumulative duration of audio recordings included in each dataset, providing an overview of the temporal distribution across different languages.

an entirely distinct collection of speakers for training and testing, eliminating the risk of data leakage and promoting model generalization. Furthermore, we not only took care of the speaker ID uniqueness but also diversified the source and location information in both the training and testing datasets. In Table 2.1, we present the sizes of training and testing datasets for each language. Additionally, It should be highlighted that considerable care was taken to guarantee class balance in both the training and testing datasets. This was achieved by ensuring that an equal number of speech segments were included in each class, which led to a fair representation of each category during the model training and evaluation phases. This data-splitting approach ensures that each dataset encapsulates a broad spectrum of linguistic variations and environmental contexts, contributing to the representativeness and richness of the training and testing data. This deliberate and comprehensive data-splitting methodology aligns with best practices, facilitating the creation of robust machine-learning models that can generalize effectively across various speakers, sources, and locations.

2.2 Audio Signal Processing and Feature Extraction

In this section, we discuss the unique approaches used to interpret the complexities of audio information: Audio Clipping, audio signal separation, and feature extraction. the audio clipping step is a process of partitioning the input audio signal into discrete, fixed-length clips. An essential technique employed for audio signal separation is Harmonic-Percussive Source

Separation (HPSS), which aims to separate complex audio signals into distinct harmonic and percussive components. Further, the feature extraction process involves using a mel spectrogram as the main feature, which comprehensively represents frequency content. These approaches give a better understanding of the analyzing audio signal and obtain the most crucial information about the sound.

2.2.1 Audio Clipping

This step embodies several key concepts related to audio signal processing. Primarily, it engages in segmentation by partitioning the input audio signal into discrete, fixed-length clips. Specifically, the process aligns with the fundamental technique of segmentation in signal processing. The function is somewhat related to windowing, although it does not strictly follow a moving window approach. It shares some characteristics with windowing, where fixed-length clips are treated as windows that are sequentially applied to the audio signal. The function implements a type of windowing where the window size is constant, and successive segments are extracted systematically. Moreover, this feature is in line with the idea of clipping. It involves dividing the input signal into shorter, predetermined segments or clips. This approach of isolating specific sections of the audio signal makes way for further analysis or processing tasks, thereby aiding the wider domain of audio signal manipulation. It is worth mentioning that the function, as it is currently implemented, does not create any overlap between consecutive clips. Instead, each clip is extracted separately, which makes it ideal for situations where non-overlapping segments are preferred. Algorithm **1** describes this approach of segmentation, windowing, and clipping the input signal into shorter, fixed-length segments.

2.2.2 Audio Signal Separation

One of the critical techniques in audio signal processing is separating audio sources, which involves decomposing a combination of audio signals into individual sources[17]. In this stage, we applied the harmonic-percussive source separation (HPSS) algorithm to split an input

Input: Input audio signal **y**, sampling rate **sr**, file name **file_name**, folder path **folder**, clip duration **clip_duration**, optional save directory **save_dir**
Output: Array of fixed-length audio clips
clips $\leftarrow$ [];
clip_length $\leftarrow$ int($clip\_duration \times sr$);
if $get\_duration(y = y, sr = sr) < clip\_duration$ **then**
    return *clips*;
end
if $get\_duration(y = y, sr = sr) < 2 \times clip\_duration$ **then**
    start_index $\leftarrow$ int($len(y)/2$) $-$ int($clip\_length/2$);
    clip $\leftarrow$ y[start_index : start_index + clip_length];
    clips.append(clip);
end
else
    n_clips $\leftarrow$ int(ceil($len(y)$/clip_length));
    for $i \leftarrow 0$ **to** $n\_clips - 1$ **do**
        start_index $\leftarrow$ $i \times$ clip_length;
        end_index $\leftarrow$ min(start_index + clip_length, len(y));
        clip $\leftarrow$ y[start_index : end_index];
        if $len(clip) == clip\_length$ **then**
            clips.append(clip);
        end
    end
end
if *save_dir is not None* **then**
    if *not exists*$(folder)$ **then**
        make_directory($folder$);
    end
    for $i \leftarrow 0$ **to** $len(clips) - 1$ **do**
        clip_path $\leftarrow$ join($folder$, f'file_name[:-4]_i.wav');
        sf.write($clip\_path$, clips[i], sr);
    end
end
return *clips*;

Algorithm 1: Pseudocode for `fixed_length_clipping` function

signal into two component signals: one comprising all harmonic sounds and the other consisting of all percussive sounds[13]. The main idea behind the HPSS algorithm is that for each audio spectrogram, harmonic sounds create horizontal patterns over time, while percussive sounds create vertical patterns in frequency[13]. Additionally, When it comes to estimating the tempo of an input signal, the percussive aspects that are extracted can be helpful indicators, and the harmonic components play an essential role in estimating chords[17]. For audio remixing, the HPSS technique can also benefit [6].

Following Section 8.1.1 of [Müller, FMP, Springer 2015][18] and (Derry FitzGerald,2010) [9], we applied harmonic-percussive separation using median filtering. Closely following the work by Fitzgerald [18], we present an HPS method as represented in Figure 2.8 The general idea is that the audio signal's spectrogram is filtered horizontally to enhance harmonic events only; conversely, the vertical direction is used to enhance percussive events only. These filtered spectrograms generate time-frequency masks that are then applied to the original spectrogram. The harmonic and percussive components of the audio signal are then obtained using an inverse Discrete Short-Time Fourier Transform(STFT) to the masked spectrogram representations. Here, we explain each step in detail:

- **First:** let x be a discrete-time representation of an audio signal as follows:

  $x : \mathbb{Z} \to \mathbb{R}$

  The goal is to decompose x into a harmonic component of the signal:

  $x^h : \mathbb{Z} \to \mathbb{R}$

  And a percussive component of the signal:

  $x^p : \mathbb{Z} \to \mathbb{R}$

  Such that :

  $x = x^h + x^p$

- **Second:** we apply the discrete Short-Time Fourier Transform(STFT) X on the audio signal x:

  $X(n,k) := \sum_{r=0}^{N-1} x(r+nH)w(r)\exp\left(-\frac{2\pi ikr}{N}\right)$

  Where N is the length size parameter, H is the hop size parameter, and $w : [0 : N-1]$ is a suitable window function of N and H.

  Then we obtain the signal's spectrogram by taking the power of function X:

  $y(n,k) := |X(n,k)|^2$

  Our decision to select $N = 2048$ and $H = 512$ was based on their critical significance, as they greatly impact the ultimate outcome of the separation process.

- **Third:** We use median filtering to create a harmonically enhanced spectrogram $\widetilde{y}_h$ and a percussively enhanced spectrogram $\widetilde{y}_p$ by filtering $\widetilde{y}$. The median is the numerical value that splits the list of numbers into two halves, with half of the numbers below it and the other half above it. To compute the median, we sort the numbers from lowest to highest and choose the middle number. In the case of an even number of observations, we usually define the median as the mean of the two middle values.

  let $A = (a_1, a_2, ..., a_L)$ be a list of length $L \in N$

  Then, the median is defined as :

$$\mu_{\frac{1}{2}}(A) : \begin{cases} a_{\frac{(L+1)}{2}} for\, L\, being\, odd \\ \frac{\left(a_{\frac{L}{2}} + a_{\frac{L}{2+1}}\right)}{2} otherwise \end{cases}$$

  Then the median filter of $L \in N$ for the list A is $\mu_{\frac{1}{2}}^{L}[A]$:

$$\mu_{\frac{1}{2}}^{L}[A](n) = u_{\frac{1}{2}}\left(\left(a_{n-\frac{(L-1)}{2}, ..., a_{n+\frac{(L-1)}{2}}}\right)\right)$$

We utilize median filtering on the spectrogram $\widetilde{y}$ in two directions: horizontally by evaluating rows of $\widetilde{y}$ and vertically by evaluating columns of $\widetilde{y}$. This produces two filtered spectrograms labeled as $\widetilde{y}^h$ and $\widetilde{y}^p$, respectively.

$$\widetilde{y}^h(n,k) := \mu_{\frac{1}{2}}\left(\left(y\left(n-\frac{(L^h-1)}{2},k\right),.....,y\left(n+\frac{(L^h-1)}{2},k\right)\right)\right)$$

$$\widetilde{y}^p(n,k) := \mu_{\frac{1}{2}}\left(\left(y\left(n,k-\frac{(L^p-1)}{2}\right),.....,y\left(n,k+\frac{(L^p-1)}{2}\right)\right)\right)$$

- **Fourth:** The two filtered spectrograms are not directly used to create the harmonic and percussive components of the audio signal. Instead, we use them to generate masks which are then used to extract the desired components from the original spectrogram. There are a variety of time-frequency masks that we can derive from $\widetilde{y}^h$ and $\widetilde{y}^p$. For this step, we utilized a binary mask where each time-frequency bin is assigned a value of either one or zero as follows:

$$M^h(n,k) := \begin{cases} 1, & if\ \widetilde{y}^h(n,k) \geq \widetilde{y}^p(n,k) \\ 0, & otherwise \end{cases}$$

$$M^p(n,k) := \begin{cases} 1, & if\ \widetilde{y}^h(n,k) < \widetilde{y}^p(n,k) \\ 0, & otherwise \end{cases}$$

We pointwise multiply the original spectrogram by each mask to obtain harmonic and percussive components. This multiplication gives us two masks: harmonic mask and percussive mask as follows:

$$y_h(n,k) := M^h(n,k)\cdot y(n,k)$$

$$y_p(n,k) := M^p(n,k)\cdot y(n,k)$$

- **Fifth:** By decomposing the spectrogram y of the audio signal into y^h and y^p, we can get two time-domain signals, x^h and x^p, by applying the two masks directly to the original Short Time Fourier Transformer(STFT) X. This leads to two complex-valued masked STFTs, X^h and X^p:

$$X_h(n,k) := M^h(n,k) \cdot X(n,k)$$

$$X_p(n,k) := M^p(n,k) \cdot X(n,k)$$

Then, applying an inverse STFT to the masked STFTs to get time-domain signal x^h and x^p:

$$x_h(r) = \sum_n \sum_k X_h(n,k) \cdot \left(w(r-nH)\exp\left(\frac{2\pi ikr}{N}\right)\right)^{-1} \cdot w(r)\exp\left(\frac{2\pi ikr}{N}\right)$$

$$x_p(r) = \sum_n \sum_k X_p(n,k) \cdot \left(w(r-nH)\exp\left(\frac{2\pi ikr}{N}\right)\right)^{-1} \cdot w(r)\exp\left(\frac{2\pi ikr}{N}\right)$$

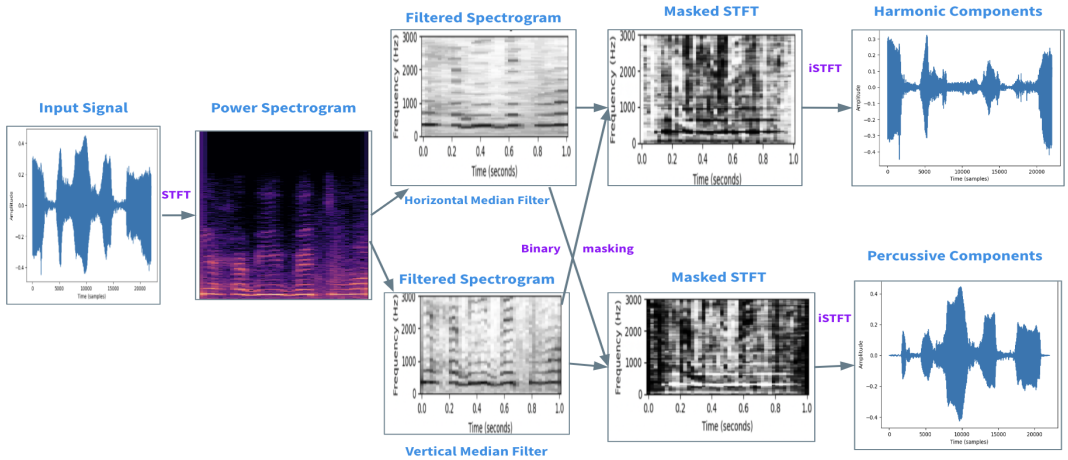


Figure 2.8. Harmonic-Percussive Source Separation (HPSS) Method

Description: This figure illustrates the Harmonic-Percussive Source Separation (HPSS) method, a signal processing technique designed to separate tonal (harmonic) and transient (percussive) components in an audio signal. The method is crucial for isolating melodies and harmonic elements from non-pitched, transient sounds that is in the infant-directed speech.

2.2.3 Feature Engineering

When extracting features from audio signal data, we use frequency-domain features like Mel spectrogram, a widely-used representation in audio processing as follows:

- We normalize the audio components, including row, harmonic, and percussive components, using a standard normalization process that scales the audio data to a typical range between 0 and 1. Subfigure 2.9a illustrates the normalized audio components.

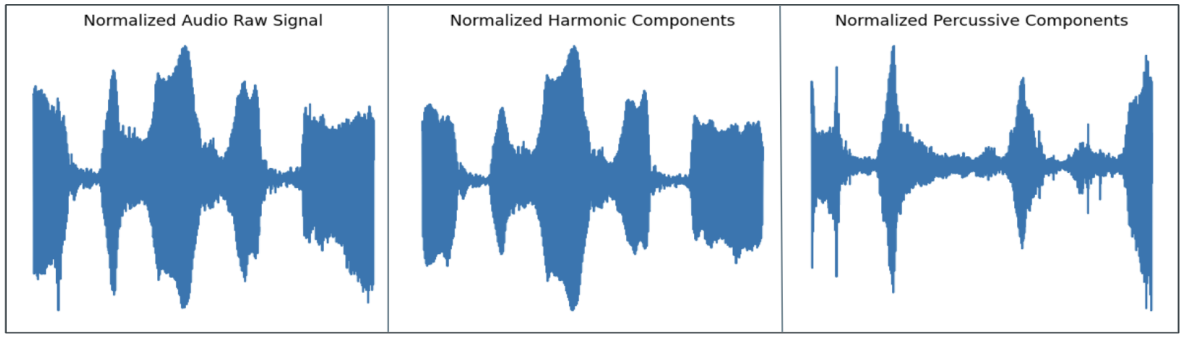
- We computed Mel spectrograms for each audio component (raw, harmonic, and percussive) using the Short-Time Fourier Transform (STFT) technique as described in the 4.1 section and the librsoa library in Python. Subfigure 2.9b shows the computed Mel spectrograms.
- We normalized each frequency bin across time frames independently using channel-wise normalization after computing the Mel spectrograms for each audio component. This process ensures that the features are not biased by extreme values in a particular frequency bin and are more robust to variations in different frequency ranges. Subfigure 2.9c displays the channel-wise normalization.
- We converted the Mel spectrograms to the decibel logarithmic scale to improve the representation and emphasize essential features. This conversion helps visualize and analyze the audio's spectral content more effectively. Subfigure 2.9d illustrates the conversion to the decibel logarithmic scale.
- The final features were obtained by stacking the normalized Mel spectrograms for each audio component (raw, harmonic, and percussive) along the third dimension, creating a multi-dimensional feature representation that preserves spectral information.

2.3 Model Development

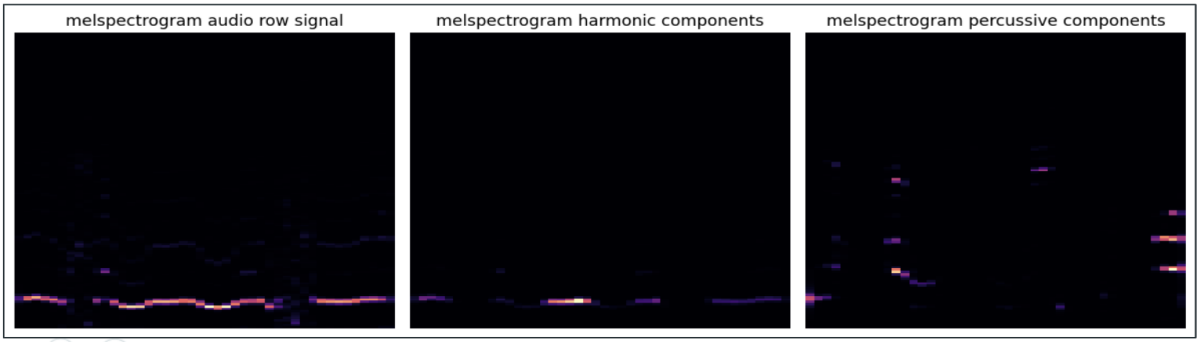
After audio signal processing and feature extraction, this section dives, in detail, into the developments of the models. Two primary model developments are in focus: **single-language model** development for each mentioned language and **multi-language model.** This section will discuss the training process and evaluation process.

2.3.1 Single-Language Model Development

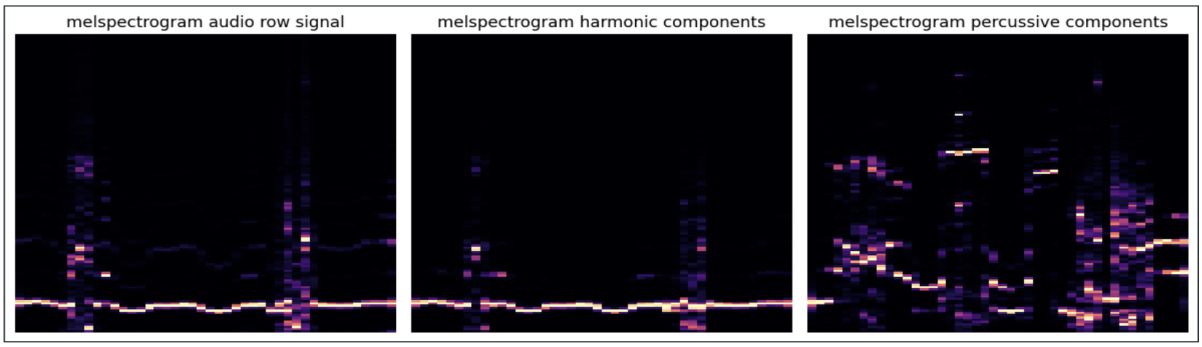
We have four datasets, one for each language: English, Arabic, Spanish, and Chinese; each training and testing dataset includes a balanced number of infant-directed and adult-directed



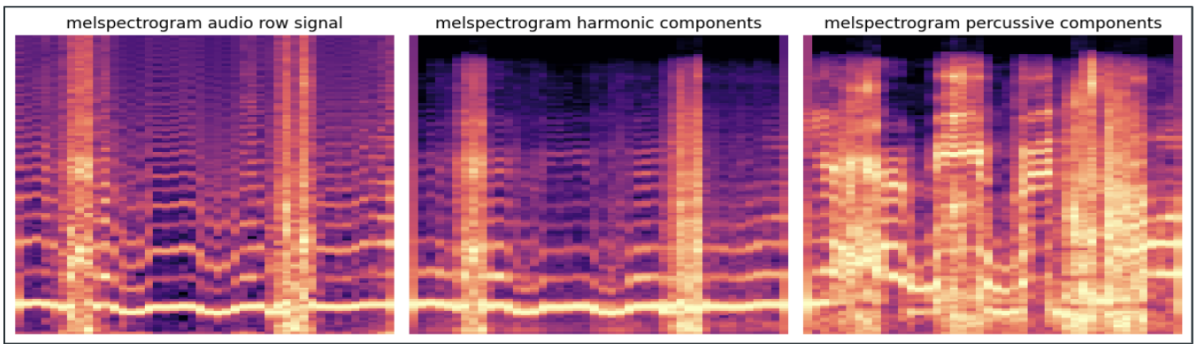
(a) Normalization of the Raw, Harmonic, and Percussive Audio Components.



(b) Mel Spectrograms for Raw, Harmonic, and Percussive Audio Component



(c) Channel-Wise Normalization for Raw, Harmonic, and Percussive Audio Component



(d) Logarithmic Scale for Raw, Harmonic, and Percussive Audio Component

Figure 2.9. Audio Feature Processing Steps

speeches. For each language, we develop a Single-Language Model using the same steps: choose the audio fixed length, training process, and evaluation process.

Fixed Length Selection:

As we work with audio datasets, we should have consent duration for all files. Initially, we set the fixed length for each audio sample to the standard duration, which is one second using the audio clipping that is discussed in Subsection 2.2.1, across all languages. This audio file length works fine with both English and Arabic single-language models, as both models successfully distinguished between Infant-directed speech(IDS) and Adult-directed speech (ADS). However, Spanish and Chinese single-language models struggled to classify IDS and ADS with a one-second length file. Consequently, to improve the Spanish single-language model's learning efficiency, we found that two-second duration audio files are better. For the Chinese single-language model, we discovered that the model did not learn with either a one-second length or two seconds, and we achieved better results when using five seconds.

Training Process:

The training process is the same across all single-language models to ensure the best possible model. The training process also includes many techniques, such as performing traditional machine learning models, neural network models, and k-fold cross-validation as hyperparameter tuning settings.

- **Shared hyperparameter tuning settings: k-fold cross-validation:** The k-fold cross-validation algorithm was used with traditional machine learning models and neural network models. According to (Jason Brownlee, 2018) [4], the general idea of using cross-validation with k-fold is to divide the training data into k subsets; in this study, we set $k = 5$. For each distinct subset, treat it as a validation set while using the remaining subsets as the training data. Next, we develop a model using the training data and assess its effectiveness on the validation dataset. Keep track of the evaluation score and disregard the model. Finally,

summarize the model's skill using a sample of the evaluation scores[4].

- **Traditional Machine Learning Models:** We use two popular traditional machine learning models: support vector machine (SVM) and Random Forest classifiers. We also performed a Grid Search with k-fold cross-validation for each classifier to find the best combination of hyperparameters. The Grid search(exhaustive search) is a Hyperparameter optimization method used to calculate the best possible values for hyperparameters[4].

  When we use the SVM algorithm, we set some hyperparameters and values for them. We specify two hyperparameters for the SVM algorithm - the regularization parameters and the kernel options. The regularization parameters, which are 0.1, 1, and 10, help balance the margin and classification error, while the kernel options include linear and radial basis functions. We systematically searched for the hyperparameter values during the grid search process to find the best combination for performance during cross-validation. Once we found the optimal hyperparameters, we set this as the best model.

  The Random Forest classifier's training process was applied using a specific hyperparameter. First is the number of estimators and decision trees in the Random Forest classifier with 100, 200, and 300 corresponding values. Second, the maximum depth of each decision tree has 10 and 20 hyperparameter values. Similar to SVM, Grid search explores the specified hyperparameter values to find the mix that performs the best during cross-validation and sets it to be the best model.

- **Neural Network Models:** In the training process of neural network models, k-fold cross-validation was applied along with other hyperparameter tuning sets, such as different model architectures, various optimizers, and distinct learning rate values. As we work with 3-channel images obtained after applying audio signal processing in Section 2.2, we chose image-based neural network model architectures. First, we use Convolutional Neural Networks (CNN), as its main application is to solve complex image-based pattern recognition tasks(O'Shea & Nash,2015)[19]. The second model's architecture is Residual

Network (ResNet) since it was used with the ImageNet dataset and achieved a low error rate of 3.57%(Kaiming He et al.,2015)[10]. Third, the pre-trained EfficientNet-B0 model was utilized as one of the model architectures. It was also trained on the ImageNet dataset reaching 93.3% accuracy(Tan & V. Le, 2020) [23]. We also employed Adam and Stochastic Gradient Descent(SGD) optimizers with one loss function: Cross-entropy. Finally, we trained all models with three learning rate values:0.01, 0.001, and 0.0001. Figure 2.10 shows the used CNN model's architecture and Figure 2.11 show the used EfficientNet-B0 model's architecture.

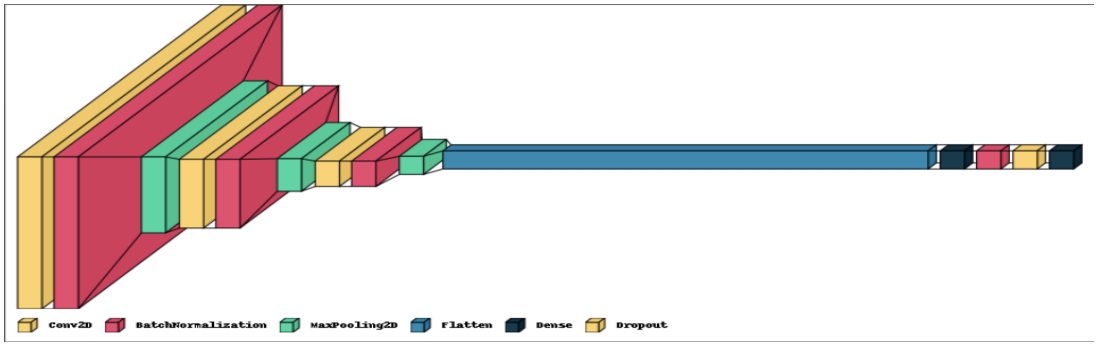


Figure 2.10. The Used CNN Model's Architecture

Description: There are two different sizes of input images: one 128x216 pixels with 3 color channels, and the second: 128x44 pixels with 3 color channels. The LeakyReLU activation function is used for introducing non-linearity in the network. Batch normalization layers are added after each convolutional and dense layer to stabilize training and improve convergence. Max pooling layers with a $(2,2)$ window size and 'same' padding are used for downsampling. A flattened layer is used in transforming 2D feature maps to a 1D feature vector. Dropout layers with a rate of 0.5 are added after dense layers to regularize the model and reduce overfitting. Then, there is the output layer with the softmax activation function for multi-class classification (in this case, 2 classes).

2.3.2 Multi-Language Model development

Inspired by the work done by (Al Rahhal et al.,2022)[2], where they built a Multilanguage Transformer for Improved Text to Remote Sensing Image Retrieval, we introduce a multilanguage model to classify infant-directed and adult-directed speech, incorporating all four different language datasets to leverage diversity and create a unified solution. Figure 2.12 shows the dataset distribution based on the language. To build such a model, we used the same approaches to make the single-language model, except for fixed-length selection.

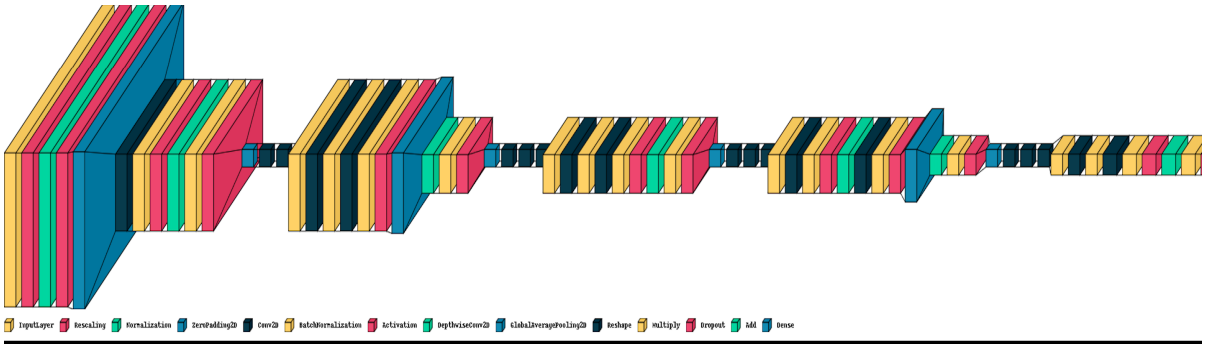


Figure 2.11. The Used EfficientNet-B0 Model's Architecture

Description: The pre-trained EfficientNetB0 model that was trained and tested on the ImageNet dataset, which is 3-channel images. Its architecture consists of multiple layers and blocks that work together to process and extract features from input images.EfficientNetB0 uses a combination of standard convolutional (layers and specialized convolutional operations like depth-wise separable convolutions.it is organized into a series of blocks, each containing a sequence of layers. These blocks are responsible for gradually extracting more complex and abstract features as the network goes deeper. The key innovation in EfficientNet is the compound scaling method, which uniformly scales the depth, width, and resolution of the network to find an optimal balance between accuracy and efficiency. It also has pooling layers to reduce the spatial dimensions of the feature maps. Fully Connected Layers that are responsible for making final predictions or producing class probabilities or scores. It also has Activation functions to introduce non-linearity into the network, including ReLU., and also has a Global Average Pooling layer to compute the average value of each feature map It helps in reducing the number of parameters and focusing on the most important features.

Fixed Length Selection:

We chose a five-second fixed duration for a multilanguage model that is trained and tested using all four datasets to ensure compatibility and accommodate the longest audio samples. This decision is based on the Chinese dataset, which requires a five-second duration for successful learning.

2.4 Evaluation Process

The evaluation process was the same for both single-language and multi-language models to measure the performance of each model and consistent across both traditional machine learning models and neural network models and it was done on the each language's testing dataset. Our approach to preparing the testing datasets involved taking specific measures to

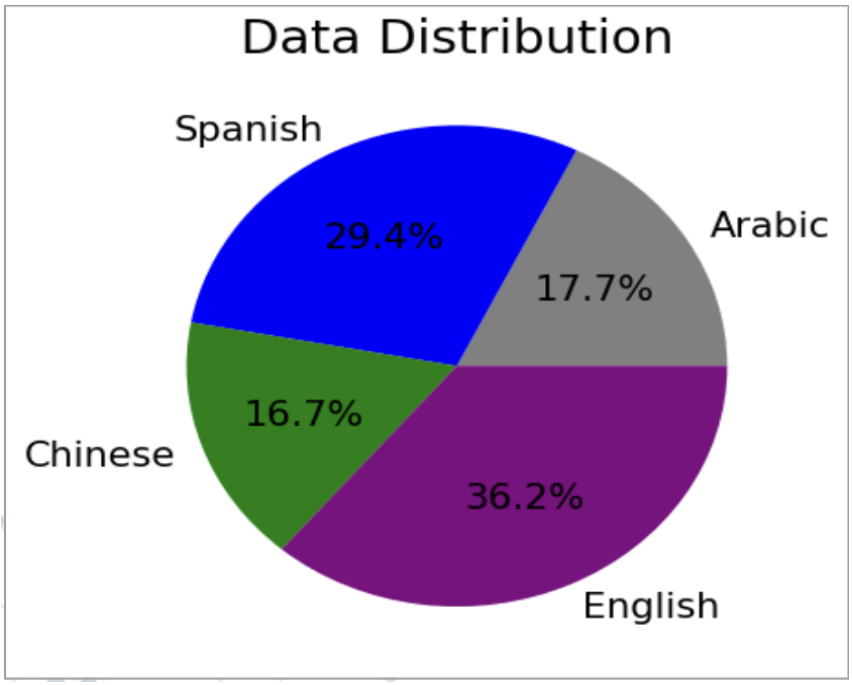


Figure 2.12. Dataset Distribution Based on the Language

Description: The figure presents a pie chart depicting the language distribution within the dataset. English accounts for 36.2%, Arabic for 17.7%, Spanish for 29.4%, and Chinese for 16.7% of the dataset. This visualization offers a clear representation of the proportion of data instances associated with each language, providing valuable insights into the linguistic diversity of the dataset.

achieve a careful class balance between the two speeches : Infant-directed speech and adult-directed speech as discussed in 2.1.5, so all testing datasets have same number of Infant-directed speech and adult-directed speech . This decision was driven by the desire to utilize a wide range of evaluation metrics, such as accuracy, recall, precision, and F1 score, to conduct a comprehensive and unbiased assessment of the model's performance. By emphasizing testing dataset class balance, we aim to provide a more nuanced and dependable evaluation process that considers the model's performance across different metrics. These metrics offer insights into different aspects of a model's predictive abilities, enabling a comprehensive assessment of its effectiveness. In addition, further analysis is performed, such as the confusion matrix and receiver operating characteristic (ROC) curve. By taking this comprehensive approach, we can gain a complete understanding of the performance of each model, thus enabling informed

decisions and optimizations across a range of applications. The following points describes each matrix's calculation:

- **Accuracy** = (Number of Correct Predictions) / (Total Number of Predictions)
- **Recall** = (True Positives) / (True Positives + False Negatives)
- **Precision** = (True Positives) / (True Positives + False Positives)
- **F1-score** = 2 * (Precision * Recall) / (Precision + Recall)

This Chapter will be submitted for publication of the material as it may appear in JMIR 2024 Papers. The thesis author was the primary investigator and author of this paper.

In the 2.1.1 section, I would like to thank Professor Lama K. Farran for her collaboration in the Arabic dataset, as 23.9% of the used Arabic dataset was shared by her. Also, I want to thank Elise A. Piazza for sharing with us his English dataset, where 9% of the used English dataset in this study was shared by him.

Chapter 3

Results and Analyses

This chapter discusses all outcomes and analyses derived from our works in developing machine-learning models that classify infant-directed and adult-directed speeches. We will be discussing two different models in separate sections. Firstly, we will examine the performance and implications of our Single-Language Model. Then, we will explore the results and cross-linguistic capabilities of our Multi-Language Model. We then engage in a thoughtful discussion, analyzing our findings to uncover patterns, challenges, and potential improvement. Additionally, a future work is discussed, where we identify unknown parts and propose for further exploration. Finally, the chapter culminates in a determined conclusion, encapsulating the essence of our research journey and charting the course for future investigations.

3.1 Single-Language Model Result

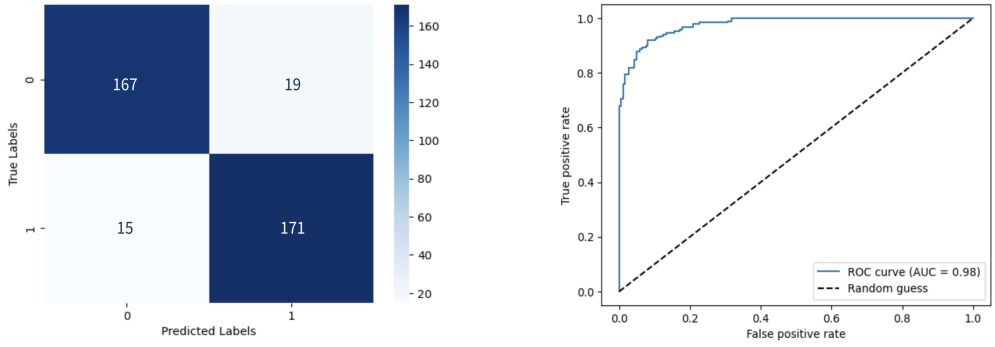
In this subsection, we provide an in-depth analysis of the performance metrics and outcomes across all models. Our comprehensive evaluation includes key metrics such as accuracy, recall, precision, and F1-score, along with a detailed examination of the confusion matrix and ROC curve for each single-language model: English, Arabic, Spanish, and Chinese models thereby providing a detailed understanding of their respective capabilities and performance.

Table 3.1. Performance Metrics of English Single-Language Model

Model	Accuracy	Recall	Precision	F1-score
SVM	0.88	0.88	0.89	0.88
RF	0.83	0.83	0.83	0.83
CNN	0.87	0.82	0.91	0.87
ResNet	0.85	0.82	0.87	0.84
Pre-trained EfficientNet-B0	**0.90**	**0.91**	**0.90**	**0.90**

English Single-Language Model Result

The result in Table 3.1 shows that the pre-trained EfficientNet-B0 model achieved the highest accuracy and f1-score as 0.90, followed by SVM and CNN, and finally, RF and ResNet achieved the lowest accuracy and F1-score. The figure 3.1a shows that the confusion matrix of the best English model performed very well and misclassified only 15 Infant-Directed speech files out of 186 files. This evidences the model's robustness in accurately discerning between different speech categories. The figure 3.1b presents the area under the curve for English model is = 0.98. These findings collectively underscore the efficacy of the pre-trained EfficientNet-B0 model in the domain of Infant-Directed speech classification, substantiating its position as the most proficient choice among the evaluated models.



(a) Confusion Matrix of the Pre-trained EfficientNet-B0 English Model

(b) ROC Curve of the Pre-trained EfficientNet-B0 English Model

Figure 3.1. Performance Evaluation of the Single-Language English Model

Description: (a) presents the confusion matrix, where the model shows great performance in classifying Infant-Directed and adult-Directed speeches and (b) illustrates the ROC curve with an overall area under the curve of 0.98.

Table 3.2. Performance Metrics of Arabic Single-Language Model

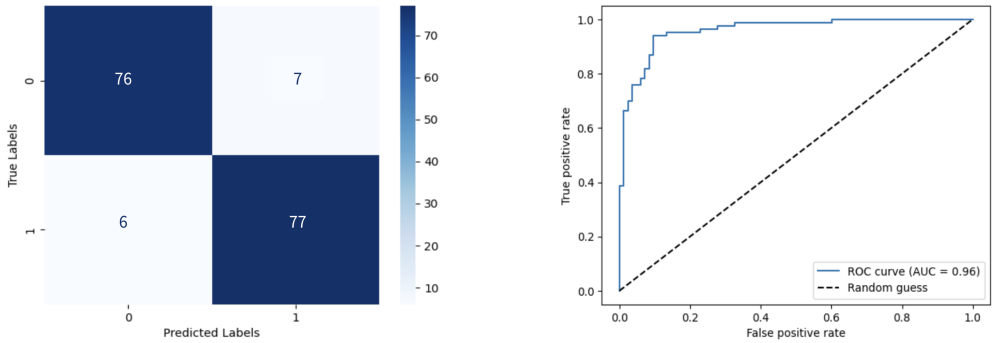
Model	Accuracy	Recall	Precision	F1-score
SVM	0.89	0.89	0.89	0.89
RF	0.90	0.90	0.90	0.90
CNN	**0.91**	**0.97**	**0.83**	**0.91**
ResNet	0.71	0.75	0.73	0.75
Pre-trained EfficientNet-B0	0.89	0.92	0.90	0.91

Arabic Single-Language Model Result

The findings from the evaluation of various models on Arabic infant-directed and adult-directed speeches classification is CNN achieved the highest scores with 0.91 f1-score and accuracy, followed by random forest and Pre-trained EfficientNet-B0 models, and again, ResNet achieved the lowest scores with 0.75 f1-score as shown in table 3.2. These results underscore the efficacy of CNN in achieving a balanced trade-off between precision and recall in this classification task. For the best Arabic model, the model performed well in identifying infant-directed speech and adult-directed speech and misclassified only 6 infant-directed Speech files out of 83 as illustrated in the sub-figure 3.2a. This showcases the model's robust discriminatory capabilities across diverse speech categories. Moreover, the ROC curve illustrated in Figure 3.2b further corroborates the model's effectiveness, boasting an area under the curve of 0.96. This metric signifies the model's capacity to distinguish between infant-directed and adult-directed speeches, affirming its reliability in Arabic model classification. In summation, the comprehensive evaluation solidifies CNN as the leading model for Arabic single-language model providing a understanding of its strengths and establishing a benchmark for future model comparisons in this domain.

Spanish Single-Language Model Result

The result of the Spanish Single-Language Model indicates, as shown in the table 3.3, that the Pre-trained EfficientNet-B0 model has the highest performance for the Spanish language model with 0.94 f1-score, and SVM has the lowest score =0.70.This stark contrast emphasizes the



(a) Confusion Matrix of the Top-Performing Arabic Model

(b) ROC Curve of the Top-Performing Arabic Model

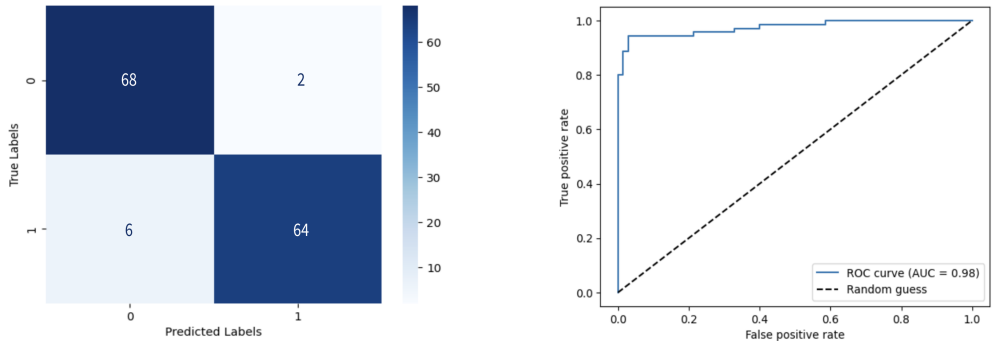
Figure 3.2. Performance Evaluation of Single-Language Arabic Model

Description: (a) illustrates the confusion matrix, and (b) depicts the ROC curve.

Table 3.3. Performance Metrics of Spanish Single-Language Model

Model	Accuracy	Recall	Precision	F1-score
SVM	0.71	0.71	0.73	0.70
RF	0.82	0.82	0.83	0.82
CNN	0.75	0.61	0.91	0.73
ResNet	0.74	0.75	0.73	0.74
Pre-trained EfficientNet-B0	**0.94**	**0.91**	**0.96**	**0.94**

superiority of the EfficientNet-B0 model in achieving a harmonious balance between precision and recall. As evidenced by the confusion matrix in sub-figure 3.3a, the best Spanish model performed well and misclassified only 6 IDS files out of 70 files. This outcome attests to the model's proficiency in distinguishing subtle nuances in infant-directed Speech and Adult-directed Speech in the Spanish language. Additionally, the ROC curve depicted in sub-figure 3.3b provides further validation of the model's reliability, showcasing an impressive area under the curve of 0.98. This metric substantiates the model's capacity to make clear distinctions between IDS and ADS, affirming its effectiveness in the classification of infant-directed speech in the Spanish language.



(a) Confusion Matrix of the Single-Language Spanish Model

(b) ROC Curve of the Single-Language Spanish Model

Figure 3.3. Performance Evaluation of the Single-Language Spanish Model

Description: (a) depicts the confusion matrix, illustrating the model's proficiency, and (b) showcases the ROC curve, emphasizing its discrimination capabilities.

Chinese Single-Language Model Result

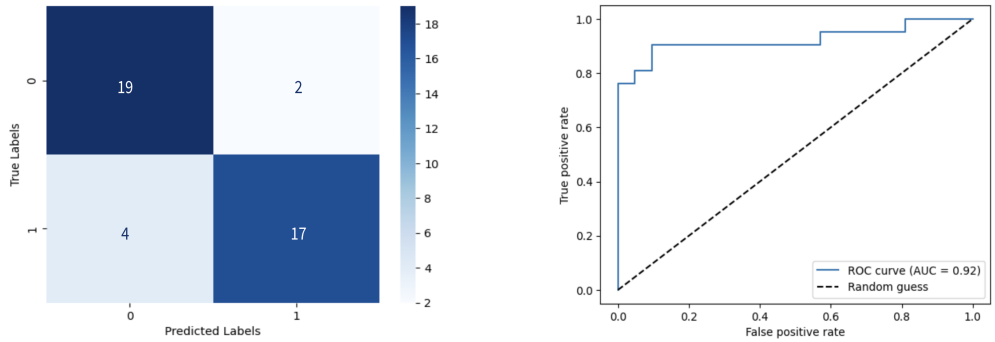
The evaluation results for the Chinese Single-Language Model, as delineated in Table 3.4 shows that the pre-trained EfficientNet-B0 model achieved the highest accuracy $= 0.88$ and f1-score $= 0.85$ with similar results from the Random forest model, and again ResNet achieved the lowest accuracy $= 0.73$. A closer examination of the confusion matrix, illustrated in Figure 3.4a, reveals the robust performance of the best Chinese single-language model where it performed well in identifying infant-directed speech and adult-directed speech and misclassified only 4 infant-directed speech files out of 21. This exemplifies the model's accuracy in capturing subtle variations in infant-directed in Chinese speech patterns. Furthermore, the ROC curve presented in Figure 3.4b underscores the model's ability to discriminate between positive and negative instances, boasting an area under the curve of 0.92.

3.2 Multi-Language Model Result

We assess the multi-language model's performance on the complete testing dataset, which includes all language datasets together. Additionally, we individually evaluate the model's performance on each language-specific testing dataset. We can see the overall performance of the

Table 3.4. Performance Metrics of Chinese Single-Language Model

Model	Accuracy	Recall	Precision	F1-score
SVM	0.76	0.76	0.76	0.76
RF	0.85	0.85	0.88	0.85
CNN	0.80	0.76	0.84	0.80
ResNet	0.73	0.90	0.67	0.77
Pre-trained EfficientNet-B0	**0.88**	**0.80**	**0.89**	**0.85**



(a) Confusion Matrix of the Top-Performing Chinese Model **(b)** ROC Curve of the Top-Performing Chinese Model

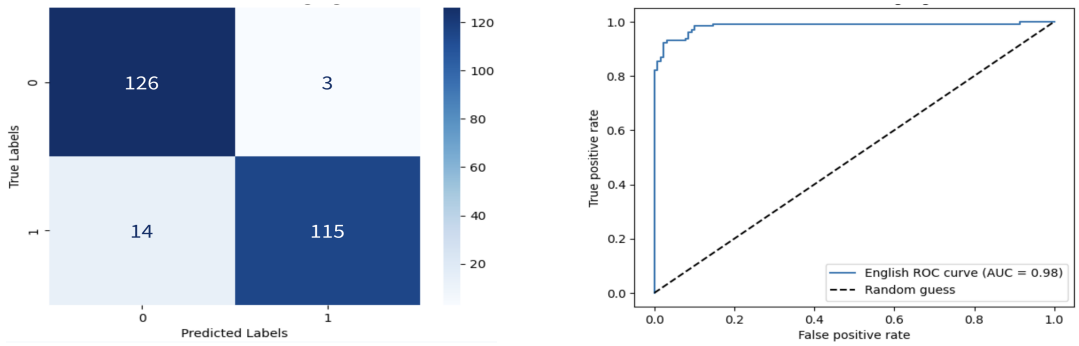
Figure 3.4. Performance Evaluation of the Single-Language Chinese Model

Description: (a) illustrates the confusion matrix, and (b) depicts the ROC curve.

multi-language model is = 0.93 accuracy and 0.89 f1-score. It also achieved a 0.96 f1-score on the English testing dataset, followed by Spanish 0.94, then Arabic 0.91, and finally Chinese 0.90. Table 3.5 provides a detailed overview of the performance metrics for the Multi-Language Model. The table includes accuracy, recall, precision, and F1-score values for each language-specific model, demonstrating the model's effectiveness in classifying different speeches in different languages. The sub-figure 3.5a shows the confusion matrix of the multi-language model, where it performs well on classifying IDS and ADS, and sub-figure 3.5b shows the overall area under the curve of multi-language model = 0.98.

Table 3.5. Performance Metrics of the Multi-Language Model

Language	Accuracy	Recall	Precision	F1-score
All language	0.93	0.92	0.86	0.89
English	0.96	0.94	1.0	0.96
Arabic	0.92	1.0	0.84	0.91
Spanish	0.93	0.89	1.0	0.94
Chinese	0.90	0.94	0.85	0.90



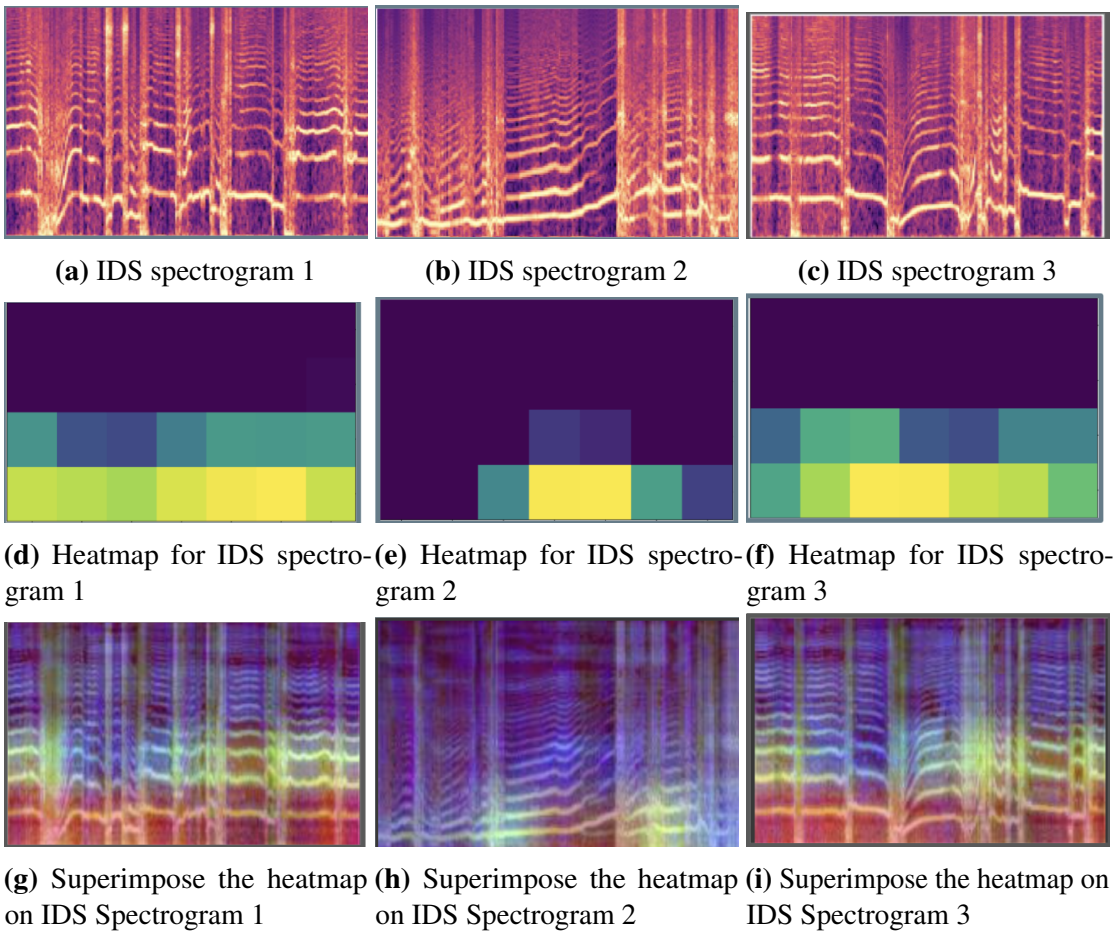
(a) Confusion Matrix of Multi-Language Model **(b)** ROC Curve of the Multi-Language Model

Figure 3.5. Multi-Language Model Results

Description: (a) illustrates the confusion matrix, where the model demonstrates excellent performance in classifying IDS and ADS, and (b) showcases the ROC curve with an overall area under the curve of 0.98.

3.3 Discussion

As we discussed earlier, one of the significant features of infant-directed speech is high pitch or high energy. Then, when we visualize its spectrogram, we are likely to see energy concentrated in specific frequency bands, as shown in figure 3.6, which correspond to the fundamental frequency and its harmonics as these harmonic components in the spectrogram represent the pitched elements of the audio. When we apply the Harmonic-Percussive Separation (HPS) algorithm to this audio, we aim to separate the audio into two components: the harmonic component and the percussive component. In this case, the harmonic component represents the pitched elements, which include the high-pitched characteristics of infant-directed speech. Meanwhile, the percussive component represents the transient and non-pitched aspects of the audio, which in the context of infant-directed speech include emotional expressiveness, intonation,



(a) IDS spectrogram 1 **(b)** IDS spectrogram 2 **(c)** IDS spectrogram 3

(d) Heatmap for IDS spectrogram 1 **(e)** Heatmap for IDS spectrogram 2 **(f)** Heatmap for IDS spectrogram 3

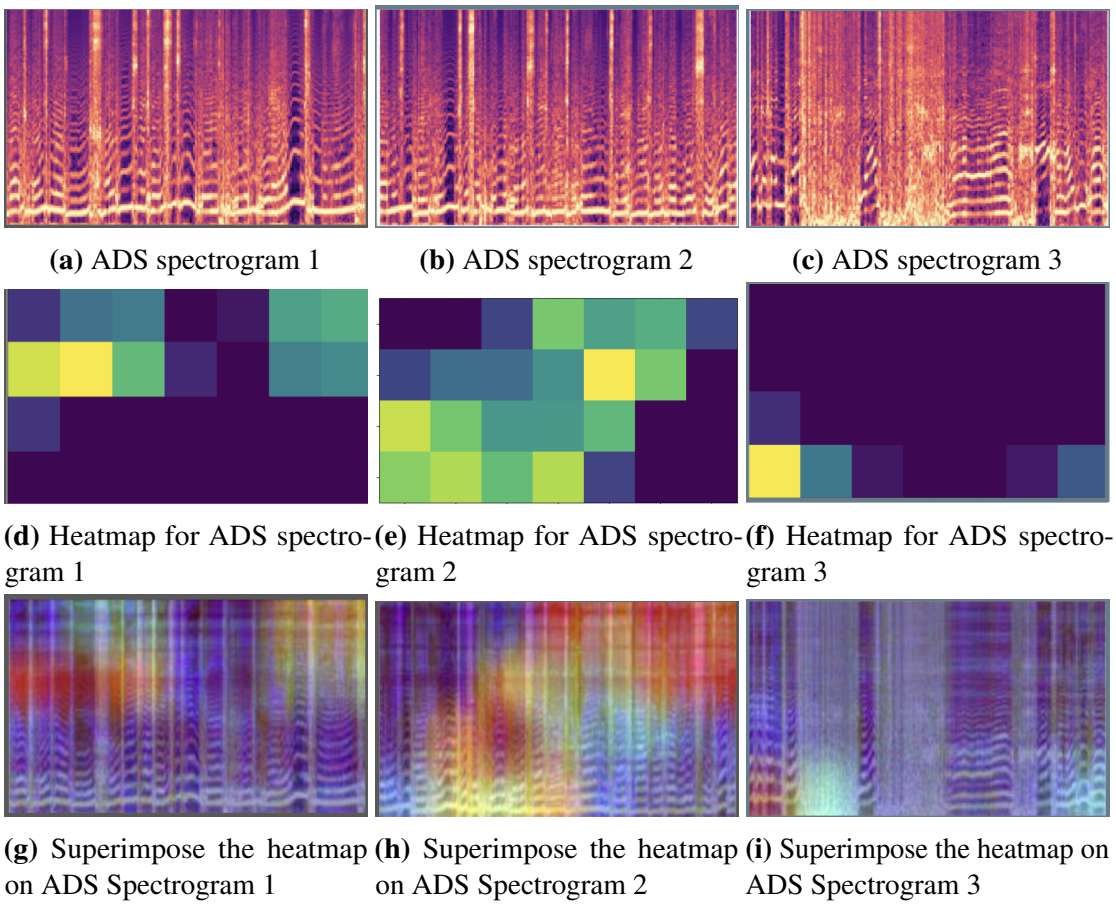
(g) Superimpose the heatmap on IDS Spectrogram 1 **(h)** Superimpose the heatmap on IDS Spectrogram 2 **(i)** Superimpose the heatmap on IDS Spectrogram 3

Figure 3.6. Infant-Directed Speech Grad-CAM

Description: The IDS Spectrogram for (a), (b), and (c). The heatmaps for (d), (e), and (f). When superimposing the heatmap and IDS spectrogram together, differences can be seen as shown in (g), (h), and (i)

and rhythmical elements. Finally, when we apply Grad-CAM to Investigate how the model is learning, we can see that for infant-directed speech, the grad-cam has a specific pattern that focuses on the harmonic components.

For adult-directed speech, when we visualize its spectrogram, the spectrogram still shows energy distributed across different frequency bands corresponding to the speech sounds. However, in adult-directed speech, the pitch is not as consistently high as in infant-directed speech. The pitch of adult-directed speech can vary depending on the context, emotion, and content of the speech. When we apply the Harmonic-Percussive Separation (HPS) algorithm to adult-directed



(a) ADS spectrogram 1 **(b)** ADS spectrogram 2 **(c)** ADS spectrogram 3

(d) Heatmap for ADS spectrogram 1 **(e)** Heatmap for ADS spectrogram 2 **(f)** Heatmap for ADS spectrogram 3

(g) Superimpose the heatmap on ADS Spectrogram 1 **(h)** Superimpose the heatmap on ADS Spectrogram 2 **(i)** Superimpose the heatmap on ADS Spectrogram 3

Figure 3.7. Adult-Directed Speech Grad-CAM

Description: The ADS Spectrogram for (a), (b), and (c). The heatmaps for (d), (e), and (f). When superimposing the heatmap and ADS spectrogram together, differences can be seen as shown in (g), (h), and (i)

speech, the resulting harmonic component still represents the pitched elements of the speech. However, in adult speech, the pitch variation is more diverse, and there might be a mix of harmonic and inharmonic (non-harmonic) components. The percussive component obtained after applying HPS represents the non-pitched elements. Finally, when we apply Grad-CAM to see how the model is learning about ADS, we can see in the figure 3.7 that the grad-cam has various patterns as the opposite of IDS grad-cams

3.4 Conclusion and Future Work

Recent estimates show a high spread of Autism Spectrum Disorder (ASD), and early identification and treatment of Autism Spectrum Disorder are essential. Additionally, recent studies show that some toddlers with Autism do not respond to high-pitched, exaggerated intonation speech, known as infant-directed speech (IDS) or motherese. Thereby, speech interactions between caregivers and young children can help detect Autism. However, studying adult-toddler interactions can be difficult due to the challenge of robustly distinguishing between Infant-Directed Speech (IDS) and Adult-Directed Speech (ADS).

As we aimed to build a model that is capable of accurately classifying IDS and ADS for multiple languages is essential, which represents the first critical step toward developing a system to study speech interactions between caregivers and young children, we achieved a significant milestone in creating a robust machine-learning algorithm capable of accurately distinguishing between Infant-Directed Speech (IDS) and Adult-Directed Speech (ADS) in four diverse languages: English, Arabic, Spanish, and Chinese. The outcomes from the single-language models emphasize the excellent accuracy in identifying IDS and ADS in Spanish and Arabic, reflected in their highest F1 scores. English and Chinese closely follow the case, displaying commendable performance across the linguistic spectrum.

Contrastingly, the multi-language model showcases superior proficiency in overall IDS and ADS detection, with English leading the performance metrics, followed by Spanish, Arabic, and Chinese. This comprehensive evaluation firms the multi-language model's superiority over individual single-language models across all languages, signifying its remarkable precision in classifying IDS and ADS.

The future work can be towards expanding our research horizons by incorporating additional languages and broadening our datasets. This expansion aims to increase the inclusively and diversity of our model's linguistic capabilities. Furthermore, we aspire to construct models focused on mother-child interaction, leveraging the insights gleaned from our research findings.

This pivotal step seeks to delve deeper into the dynamics of specific speech interactions, opening new avenues for understanding and enhancing communication dynamics within varied linguistic contexts.

Some content of this Chapter will be submitted for publication of the material as it may appear in JMIR 2024 Papers. The thesis author was the primary investigator and author of this paper.

Bibliography

[1] Najla D Al Futaisi, Alejandrina Cristia, and Björn W Schuller. Hearttoheart: The arts of infant versus adult-directed speech classification. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[2] Mohamad M Al Rahhal, Yakoub Bazi, Norah A Alsharif, Laila Bashmal, Naif Alajlan, and Farid Melgani. Multilanguage transformer for improved text to remote sensing image retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:9115–9126, 2022.

[3] American Psychiatric Association. Home, Year.

[4] Jason Brownlee. A gentle introduction to k-fold cross-validation, May 23 2018.

[5] Centers for Disease Control and Prevention. Facts about autism spectrum disorders, Year.

[6] Christian Dittmar and Meinard Müller. Reverse engineering the amen break—score-informed separation and restoration applied to drum recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1535–1547, 2016.

[7] Lama K Farran, Chia-Cheng Lee, Hyunjoo Yoo, and D Kimbrough Oller. Cross-cultural register differences in infant-directed speech: An initial study. *PloS one*, 11(3):e0151518, 2016.

[8] Marisa G Filipe, Linda Watson, Selene G Vicente, and Sónia Frota. Atypical preference for infant-directed speech as an early marker of autism spectrum disorders? a literature review and directions for further research. *Clinical Linguistics & Phonetics*, 32(3):213–231, 2018.

[9] Derry FitzGerald. Harmonic/percussive separation using median filtering. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 246–253, Graz, Austria, 2010.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Bertolo M Lee-Rubin H Amir D Bainbridge CM Simson J Knox D Glowacki L Alemu E Galbarczyk A Jasienska G Ross CT Neff MB Martin A Cirelli LK Trehub SE Song J

Kim M Schachner A Vardy TA Atkinson QD Salenius A Andelin J Antfolk J Madhivanan P Siddaiah A Placek CD Salali GD Keestra S Singh M Collins SA Patton JQ Scaff C Stieglitz J Cutipa SC Moya C Sagar RR Anyawire M Mabulla A Wood BM Krasnow MM Mehr SA. Hilton CB, Moser CJ. Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*, 6(11):1545–1556, 2022.

[12] Takao Inoue, Ryuta Nakagawa, Misa Kondou, Tadashi Koga, and Kazuyuki Shinohara. Discrimination between mothers' infant-and adult-directed speech using hidden markov models. *Neuroscience research*, 70(1):62–70, 2011.

[13] Thomas Pr¨atzlich Jonathan Driedger. Harmonic percussive source separation. https://www.audiolabs-erlangen.de/content/05-fau/professor/00-mueller/02-teaching/2016w_mpa/LabCourse_HPSS.pdf, 2016.

[14] Patricia K Kuhl, Jean E Andruski, Inna A Chistovich, Ludmilla A Chistovich, Elena V Kozhevnikova, Viktoria L Ryskina, Elvira I Stolyarova, Ulla Sundberg, and Francisco Lacerda. Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5326):684–686, 1997.

[15] Jialu Li, Mark Hasegawa-Johnson, and Nancy L McElwain. Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations. *Speech communication*, 133:41–61, 2021.

[16] Weiyi Ma, Roberta Michnick Golinkoff, Derek M Houston, and Kathy Hirsh-Pasek. Word learning in infant-and adult-directed speech. *Language Learning and Development*, 7(3):185–201, 2011.

[17] Yoshiki Masuyama, Kohei Yatabe, and Yasuhiro Oikawa. Phase-aware harmonic/percussive source separation via convex optimization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 985–989. IEEE, 2019.

[18] Meinard Müller. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[19] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[20] Erika Parlato-Oliveira, Mohamed Chetouani, Jean-Maximilien Cadic, Sylvie Viaux, Zeineb Ghattassi, Jean Xavier, Lisa Ouss, Ruth Feldman, Filippo Muratori, and Catherine Saint-Georges Cohen, David Cohen. The emotional component of infant directed-speech: a cross-cultural study using machine learning. *Neuropsychiatrie de l'Enfance et de l'Adolescence*, 68(2):106–113, 2020.

[21] Elise A Piazza, Marius Cătălin Iordan, and Casey Lew-Williams. Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology*, 27(20):3162–3167, 2017.

[22] Karen Pierce, Teresa H Wen, Javad Zahiri, Charlene Andreason, Eric Courchesne, Cynthia C Barnes, Linda Lopez, Steven J Arias, Ahtziry Esquivel, and Amanda Cheng. Level of attention to motherese speech as an early marker of autism spectrum disorder. *JAMA network open*, 6(2):e2255125–e2255125, 2023.

[23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[24] Yaqiong Xiao, Teresa H Wen, Lauren Kupis, Lisa T Eyler, Disha Goel, Keith Vaux, Michael V Lombardo, Nathan E Lewis, Karen Pierce, and Eric Courchesne. Neural responses to affective speech, including motherese, map onto clinical and social eye tracking profiles in toddlers with asd. *Nature Human Behaviour*, 6(3):443–454, 2022.