Role of acoustics and prosodic features for children's age classification

Vishakha Kumari, Abhijit Sinha and Hemant Kumar Kathania
Department of Electronics and Communication Engineering, NIT Sikkim, India
b2100121@nitsikkim.ac.in, phec230023@nitsikkim.ac.in and hemant.ece@nitsikkim.ac.in

2024 International Conference on Signal Processing and Communications (SPCOM) | 979-8-3503-5045-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/SPCOM60851.2024.10631624

***Abstract*—A speech signal contains numerous details about the speaker, including their gender, age, accent, and health. In recent times, deep neural networks have shown remarkable abilities in identifying the gender and age of human speech signals. Still, not much research has been done on using speech signals to determine a child's age. In this paper, we have explored the role of acoustics and prosodic features in children's age classification. Further, 23 unique combinations of acoustics and prosody features were also explored to boost the classifier's performance. The databases utilized for the age classification experiments are the PF-Star and CMU Kids children's speech corpus. For the PF-Star dataset, the feature combination of MFCC and Tempo, demonstrated the best classification accuracy of 88.64%. Whereas our best feature combination, containing MFCC, Chroma and Spectral features, achieved the best classification accuracy of 94.73% using an ANN classifier. The aforementioned unique feature combinations show an absolute improvement of 4.55% and 1.46%, in classification of children's age for PF-Star and CMU Kids respectively, over the baseline performances.**

***Index Terms*—Children age classification, ANN, PCA, prosodic features, acoustic features**

I. Introduction

Speech is a complex phenomenon whose production involves the movements of numerous anatomical structures that affect voice characteristics and overall speech quality [1]. As the primary and most convenient form of communication, speech also encompasses speaker-dependent para-linguistic data in addition to linguistic information, such as the speaker's identity, emotional state, physical state, age, or gender [2]. The availability of speech analysis and synthesis algorithms has spurred extensive research in the field of speech processing, encompassing tasks like speech recognition, speaker verification, and identity recognition. However, an area that has been relatively unexplored is age classification, specifically designed to tackle the distinct characteristics and challenges presented by children [3].

Age classification for children is a critical factor impacting education, healthcare, and various sectors [4]. It guides the development of age-appropriate curricula, aids in medical treatments tailored to specific age groups, and informs research on child development. From child protection services to marketing and legal considerations, age classification plays a crucial role in ensuring the well-being of children across different domains [2]. In this context, understanding and accurately classifying children's ages holds profound implications across diverse domains, ranging from educational settings to online safety measures. As children traverse through different stages of physical and cognitive development, discerning their age becomes a critical task with far-reaching consequences [5]. Whether in the context of age-appropriate content delivery, personalized learning experiences, or ensuring child safety in digital spaces, the ability to automatically and precisely classify children's ages has become unavoidable for addressing the unique characteristics and challenges posed by children.

As voices naturally undergo changes over time, the intricate task of capturing acoustic nuances that delineate different stages of childhood becomes both specific and complex. Numerous researchers have delved into the field of age classification, exploring a range of Machine Learning (ML) and Deep Learning (DL) approaches [6]–[8]. These investigations have showcased the utilization of diverse classifiers to tackle age classification tasks [9]–[11], including efforts to optimize accuracy through state-of-the-art frameworks. Despite these advancements, the exploration of acoustics and prosody features for children's age classification remains limited.

This study aims to fill this gap by thoroughly exploring various acoustic and prosodic features specifically tailored for children's age classification. To select the baseline model, we explored four different classifiers, i.e., Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Random Forest (RF), and Artificial Neural Network (ANN), with MFCC as input features. Then we assessed the impact of 23 unique feature set combinations on the best classifier among the four. Our approach aims to capture distinctive characteristics in speech that correlate with age. By exploring both acoustic and prosodic elements [12], [13], we seek to enhance the precision of age classification in children. This paper contributes to the broader understanding of utilizing speech features for age categorization, providing insights for potential applications in diverse fields such as education, healthcare, and speech technology.

II. Corpus Details

We have used two distinct children speech datasets, PF-Star [14] and CMU Kids [15], to comprehensively explore and train our models for Children Age Classification through speech.

A. PF-STAR

The PF-STAR British English children's speech corpus [14] was compiled at the University of Birmingham, UK. This dataset comprises recordings from 158 children aged 4 to 14, encompassing diverse speech materials recorded at university laboratories and primary schools. Materials include "SCRIBE" sentences, isolated words, phonetically rich sentences, generic phrases, an 'accent diagnostic' passage ('sailor passage'), and digit triples.

B. The CMU Kids Corpus

The CMU Kids [15] dataset consists of sentences read aloud by children. The dataset includes recordings from children aged 6 to 11 comprising 24 male and 52 female speakers, totaling 76 speakers. In total, the dataset consists of 5,180 utterances. The primary aim of this dataset is to facilitate research in automatic speech recognition, particularly focusing on children's speech patterns.

III. Explored Acoustics and Prosodic features

A. Mel-Frequency Cepstral Coefficients (MFCC)

MFCC, a widely employed feature in speech and audio signal processing, forms the foundation of our feature extraction process [16]. It has been used to many different areas, particularly audio signal processing, where it is utilised for speaker identification, voice recognition, and gender identification [17], [18]. Through a multi-step procedure involving framing, windowing, spectral analysis, and transformation, MFCCs offer a compact yet informative representation of the spectral content of speech.

B. Chroma Features

Chroma features [19] in speech analysis depict energy distribution across pitch classes, offering insights into harmonic content. The mean of chroma features is widely used to succinctly capture a speech signal's tonal characteristics, providing a condensed representation of pitch class distribution.

C. Formants

Formants [3] are resonant frequencies in the acoustic spectrum of speech sounds, particularly vowels, representing the peaks in the spectral envelope. They provide information about the vocal tract's resonant properties, which can change with age. As individuals age, physiological changes in the vocal tract affect the distribution of formant frequencies. Analyzing formants allows for capturing age-related variations in speech sounds.

D. Tonnetz

Tonnetz features [20] in speech processing involve mapping the pitch classes of audio signals onto a three-dimensional space using mathematical transformations. This geometric representation captures tonal relationships and is particularly useful for analyzing the tonal structure of speech signals. Tonnetz features provide a compact representation of pitch information, aiding tasks such as intonation analysis and prosody recognition in speech processing applications.

E. Pitch

Pitch-related features [21], [22] are used for speech processing tasks because they capture variations in vocal characteristics. Changes in pitch patterns over the life course, such as the gradual deepening of voices with age, are reflected in these features. Analyzing the vocal range and central tendency of pitch distribution helps machine learning models discern age-related patterns.

F. Log Energy, Delta Log Energy, Delta Delta Log Energy

Log energy represents the logarithm of the energy of a signal and provides a robust measure of its intensity. Delta log energy captures the rate of change of log energy over time, while delta delta log energy represents the acceleration of this change. These features are employed because they encapsulate dynamic aspects of speech signals, reflecting variations in vocal effort, fluency, and expressiveness.

G. Spectral Features

Spectral centroid frequency, Spectral centroid magnitude, Spectral Rolloff, Spectral Bandwidth, Spectral Flatness, and Spectral Contrast are features derived from the spectral content of audio signals, providing insights into frequency distribution and tonal qualities [23]–[25]. In voice-based age classification, these features are valuable because age-related changes in vocal characteristics, such as pitch and spectral content, can be reflected in these measures. Extracting and analyzing these spectral features offer a comprehensive representation of voice characteristics, aiding in identifying age-related patterns for effective classification.

H. Root Mean Square (RMS)

Root Mean Square (RMS) is a statistical measure that calculates the average amplitude of a signal by taking the square root of the mean of the squared values of the signal samples. RMS is employed as a feature extraction method to capture the energy or loudness of a vocal signal. By utilizing RMS as a feature, one can quantify the overall energy level of a voice, which may exhibit patterns associated with different age groups.

I. Short-Term Energy and Zero Crossing Rate

Short-Term Energy and Zero Crossing Rate [26], [27] are audio signal features widely used in various audio processing applications. Short-Term Energy measures the signal's energy within short, consecutive frames, providing insights into the signal's intensity. Zero Crossing Rate, on the other hand, quantifies how often the audio waveform crosses the zero-axis within a given time frame, offering information about the signal's pitch and noisiness.

J. Mean, Std & Median Amplitude

Mean Amplitude represents the average strength of a voice signal, Std Amplitude indicates its variability, and Median Amplitude reflects the middle point of amplitude distribution. These features capture age-related changes in vocal strength and stability. They serve as valuable indicators of developmental variations in the vocal apparatus.

K. Tempo

Tempo [28] (speech rate) represents the speed or pace of a musical piece, while Mean Tempo is the average tempo over a duration. Tempo features can capture age-related variations in speech rhythm and pacing, reflecting developmental changes. Children often exhibit faster tempos, while adults may have a more measured pace.

L. Onset Frames

Onset frames represent the points in an audio signal where sound begins. The timing and regularity of speech onsets may vary with age, reflecting developmental changes in vocal patterns. Analyzing the timing of speech onset can reveal how individuals initiate vocalizations.

M. Harmonic-to-Noise Ratio (HNR)

Harmonic-to-Noise Ratio (HNR) [29] is a measure indicating the balance between harmonically structured components and noise in a voice signal. HNR reflects changes in the vocal tract related to age. Children typically exhibit higher harmonic content due to smaller vocal tract dimensions, making HNR an effective feature for capturing age-specific vocal characteristics.

N. Skewness and Kurtosis

In voice-based age classification, skewness and kurtosis [5] serve as essential statistical measures. Skewness gauges the asymmetry of feature distribution, revealing biases towards specific age groups. Meanwhile, kurtosis assesses the peakedness and tail characteristics of the distribution, providing insights into feature concentration and outlier presence. Analyzing skewness and kurtosis collectively enhances our understanding of the statistical properties of voice data, contributing to the refinement of age classification models for increased accuracy and robustness.

O. Shannon Entropy

Shannon Entropy [30] measures the information content or uncertainty in a signal. Children's voices undergo distinct developmental changes, and higher entropy values may indicate more diverse and less predictable pitch patterns, contributing valuable information for age classification models. Each feature has its own significance in capturing different aspects of the audio signal.

IV. Proposed Feature Combinations

In our study on voice-based age classification, we utilized MFCC as the base feature. We systematically combined MFCCs with various features to assess their impact on classification performance. Table I provides 23 different combinations of feature sets explored in our study. This approach aimed to create a comprehensive feature set that collectively represents diverse acoustic and prosodic properties associated with age-related vocal changes. In Table I we have also given total number of features after combinations and set number for each combination with MFCC feature for further use .

TABLE I
THIS TABLE DESCRIBES THE 23 FEATURE SET COMBINATIONS WITH THE BASE FEATURE MFCC EXPLORED IN OUR STUDY

S.No.	Feature Combinations	No. of Features	Set No.
1.	MFCC + Chroma	38	set 1
2.	MFCC + Formant	31	set 2
3.	MFCC + Log Energy + Delta Log Energy + Delta Delta Log Energy	29	set 3
4.	MFCC + Pitch	29	set 4
5.	MFCC + Tonnetz	32	set 5
6.	MFCC + Amplitude	29	set 6
7.	MFCC + Tempo	28	set 7
8.	MFCC + Onset Frames	27	set 8
9.	MFCC + Spectral	39	set 9
10.	MFCC + Harmonic to Noise Ratio	27	set 10
11.	MFCC + Skewness	27	set 11
12.	MFCC + RMS Values	27	set 12
13.	MFCC + Kurtosis	27	set 13
14.	MFCC + Short Term Energy	27	set 14
15.	MFCC + Shannon Entropy	27	set 15
16.	MFCC + Chroma + Spectral	51	set 16
17.	MFCC + Spectral + ZCR + RMS	41	set 17
18.	MFCC + Chroma + Spectral + ZCR + RMS	53	set 18
19.	MFCC + Chroma + Spectral + Pitch + Log Energy	53	set 19
20.	MFCC + Amplitude + Tempo + Onset Frames + Harmonic to Noise Ratio + Skewness + Kurtosis + Short Term Energy + Shannon Entropy	37	set 20
21.	MFCC + ZCR Values	27	set 21
22.	MFCC + All	81	set 22
23.	MFCC + Chroma + Pitch Values + Spectral Rolloff + Spectral Bandwidth + Spectral Flatness + Spectral Contrast + Median Spectral Bandwidth + Median Spectral Centroid + Std Spectral Contrast	46	set 23

V. Results and Discussions

A. Experimental setup and Baseline

Initially with the exploration of four different classifiers we opted a suitable classifier based on its evaluation on the base MFCC feature set. Table II presents the baseline scores for different classifiers— Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), and Artificial Neural Network (ANN) utilizing a baseline feature set of 26 MFCC coefficients. The evaluation includes Accuracy (A), Precision (P), Recall (R), and F1 Score (F1) on two datasets: PF STAR and CMU kids. From all four classifiers, ANN gives the best performance compared to others.

TABLE II
THIS TABLE DESCRIBES THE COMPARATIVE PERFORMANCE FOR A RANGE OF CLASSIFIERS—SVM, KNN, RANDOM FOREST, AND ANN, ON THE PF-STAR AND CMU KIDS DATASETS. THE TABLE PRESENTS KEY PERFORMANCE METRICS, INCLUDING ACCURACY (A) IN PERCENTAGE, PRECISION (P), RECALL (R), AND F1 SCORE (F1).

Classifiers	PF STAR				CMU			
	A	P	R	F1	A	P	R	F1
SVM	84.01	0.83	0.84	0.83	87.8	0.89	0.88	0.88
KNN	81.82	0.82	0.82	0.8	87.99	0.89	0.88	0.88
Random Forest	80.3	0.8	0.8	0.79	78.59	0.81	0.79	0.78
ANN	**84.09**	0.85	0.84	0.83	**93.27**	0.93	0.93	0.93

B. Effect of feature combination and PCA

From the analysis in Section V-A we found ANN to be best suited classifier for our task in hand. The subsequent set of experiments were performed utilising the ANN classifier. ANN's ability to balance precision, recall, and overall accuracy makes it a robust choice for our subsequent analyses and reinforces its suitability for age classification tasks. Table III details the performance on all the combination sets of features enlisted in Table I. For PF-Star feature set 23 and feature set 16 for CMU Kids give best performance compared to other features combinations for children age classification.

Furthermore, we applied Principal Component Analysis (PCA) to the feature sets to manage potential dimensionality issues and improve computational efficiency. PCA helps capture the most essential information while reducing the overall dimensionality of the feature space. In our study we considered

TABLE III

THIS TABLE DESCRIBES THE PERFORMANCE METRICS OF ANN CLASSIFIER ON PF-STAR AND CMU KIDS DATASETS WITH VARIOUS FEATURE SET COMBINATIONS. THE TABLE PRESENTS KEY PERFORMANCE METRICS, INCLUDING ACCURACY (A) IN PERCENTAGE, PRECISION (P), RECALL (R), AND F1 SCORE (F1).

Feature Set	PF-Star				CMU Kids			
	A	P	R	F1	A	P	R	F1
MFCC (Baseline)	84.09	0.85	0.84	0.83	93.27	0.93	0.93	0.93
set 1	**85.61**	0.88	0.86	0.84	**94.35**	0.94	0.94	0.94
set 2	82.58	0.83	0.83	0.81	92.82	0.93	0.93	0.93
set 3	**84.85**	0.86	0.85	0.84	**93.46**	0.94	0.94	0.94
set 4	**85.61**	0.87	0.87	0.84	92.63	0.93	0.93	0.93
set 5	81.82	0.83	0.82	0.81	92.19	0.92	0.92	0.92
set 6	**85.61**	0.87	0.86	0.84	**93.96**	0.94	0.94	0.94
set 7	**85.61**	0.87	0.86	0.84	92.19	0.92	0.92	0.92
set 8	84.09	0.86	0.84	0.83	**93.39**	0.93	0.93	0.93
set 9	82.58	0.84	0.83	0.81	**93.33**	0.93	0.93	0.93
set 10	**85.61**	0.87	0.86	0.84	**93.46**	0.93	0.93	0.93
set 11	**84.85**	0.86	0.85	0.84	**93.33**	0.93	0.93	0.93
set 12	**84.85**	0.86	0.85	0.83	**93.84**	0.94	0.94	0.94
set 13	**84.85**	0.86	0.85	0.83	92.63	0.93	0.93	0.93
set 14	**84.85**	0.86	0.85	0.83	**93.39**	0.93	0.93	0.93
set 15	**86.36**	0.87	0.86	0.85	92.82	0.93	0.93	0.93
set 16	**84.85**	0.86	0.85	0.84	**94.33**	0.95	0.95	0.95
set 17	81.82	0.83	0.82	0.8	**93.71**	0.94	0.94	0.94
set 18	**85.61**	0.88	0.86	0.84	**94.41**	0.94	0.94	0.94
set 19	**84.85**	0.87	0.85	0.84	**93.33**	0.93	0.93	0.93
set 20	**85.61**	0.87	0.86	0.85	93.01	0.93	0.93	0.93
set 21	**85.61**	0.87	0.86	0.84	**93.33**	0.93	0.93	0.93
set 22	**84.85**	0.86	0.85	0.84	92.5	0.93	0.93	0.93
set 23	**86.36**	0.89	0.86	0.85	**93.96**	0.94	0.94	0.94

TABLE IV

THIS TABLE DESCRIBES THE PERFORMANCE METRICS OF ANN CLASSIFIER ON PF-STAR AND CMU KIDS DATASET WITH VARIOUS FEATURE SET COMBINATIONS FOR PCA VARIANCE RATIO 0.98. THE TABLE PRESENTS KEY PERFORMANCE METRICS, INCLUDING ACCURACY (A) IN PERCENTAGE, PRECISION (P), RECALL (R), AND F1 SCORE (F1).

Feature Set	PF-Star				CMU Kids			
	A	P	R	F1	A	P	R	F1
MFCC (Baseline)	84.09	0.85	0.84	0.83	93.27	0.93	0.93	0.93
set 1	**88.64**	0.90	0.88	0.87	**94.09**	0.94	0.94	0.94
set 2	84.09	0.85	0.84	0.83	93.27	0.93	0.93	0.93
set 3	**88.36**	0.85	0.83	0.82	**93.52**	0.94	0.94	0.94
set 4	**86.36**	0.87	0.86	0.85	92.44	0.92	0.92	0.92
set 5	**87.88**	0.89	0.88	0.87	91.68	0.92	0.92	0.92
set 6	**85.61**	0.87	0.86	0.84	**93.84**	0.94	0.94	0.94
set 7	**87.88**	0.90	0.88	0.87	91.99	0.92	0.92	0.92
set 8	**86.36**	0.89	0.86	0.85	93.2	0.93	0.93	0.93
set 9	83.33	0.86	0.83	0.82	92.82	0.93	0.93	0.93
set 10	**86.36**	0.88	0.86	0.85	91.93	0.92	0.92	0.92
set 11	**84.85**	0.86	0.85	0.84	**93.58**	0.94	0.94	0.94
set 12	**85.61**	0.86	0.86	0.84	92.88	0.93	0.93	0.93
set 13	**84.85**	0.86	0.85	0.84	92.19	0.92	0.92	0.92
set 14	**85.61**	0.86	0.86	0.84	92.88	0.93	0.93	0.93
set 15	**87.88**	0.89	0.88	0.87	**92.63**	0.93	0.93	0.93
set 16	**84.85**	0.87	0.85	0.84	**93.33**	0.93	0.93	0.93
set 17	84.09	0.86	0.84	0.83	**93.52**	0.94	0.94	0.94
set 18	**85.61**	0.88	0.86	0.84	**93.9**	0.94	0.94	0.94
set 19	**86.36**	0.88	0.86	0.85	**94.35**	0.94	0.94	0.94
set 20	**86.36**	0.87	0.86	0.85	92.95	0.93	0.93	0.93
set 21	83.33	0.85	0.83	0.82	**93.58**	0.94	0.94	0.94
set 22	**85.61**	0.87	0.86	0.84	93.14	0.93	0.93	0.93
set 23	**84.85**	0.87	0.85	0.84	**94.35**	0.94	0.94	0.94

three variance ratios (0.98, 0.96, 0.94) during PCA. This analysis aimed to find the optimal balance between feature expressiveness and computational efficiency, investigating how dimensionality reduction affects classification accuracy.

It can be clearly observed from the experimental results in Table IV,V and VI that most of feature combinations showed improved accuracy compared to the baseline. Introducing PCA with a level of 0.98 in Table IV demonstrates improved accuracy for PF-Star, with feature Set 2 reaching 88.64%. While for CMU Kids these variations are not much impacting on the overall classification accuracy. For PCA factor 0.96 in Table VI feature set 19 is giving the best accuracy of 94.79% for CMU Kids.

TABLE V

THIS TABLE DESCRIBES THE PERFORMANCE METRICS OF ANN CLASSIFIER ON PF-STAR AND CMU KIDS DATASET WITH VARIOUS FEATURE SET COMBINATIONS FOR PCA VARIANCE RATIO 0.96. THE TABLE PRESENTS KEY PERFORMANCE METRICS, INCLUDING ACCURACY (A) IN PERCENTAGE, PRECISION (P), RECALL (R), AND F1 SCORE (F1).

Feature Set	PF-Star				CMU Kids			
	A	P	R	F1	A	P	R	F1
MFCC (Baseline)	84.09	0.85	0.84	0.83	93.27	0.93	0.93	0.93
set 1	**85.61**	0.87	0.86	0.84	**94.16**	0.94	0.94	0.94
set 2	84.09	0.85	0.84	0.82	92.82	0.93	0.93	0.93
set 3	**86.36**	0.87	0.86	0.85	**93.39**	0.93	0.93	0.93
set 4	**85.61**	0.87	0.86	0.84	92.12	0.92	0.92	0.92
set 5	**85.61**	0.87	0.86	0.85	90.47	0.91	0.90	0.90
set 6	**85.61**	0.87	0.86	0.84	93.27	0.93	0.93	0.93
set 7	**88.64**	0.90	0.89	0.88	92.19	0.92	0.92	0.92
set 8	**86.36**	0.88	0.86	0.85	92.5	0.93	0.93	0.93
set 9	**83.33**	0.86	0.83	0.82	92.95	0.93	0.93	0.93
set 10	**86.36**	0.88	0.86	0.85	91.36	0.91	0.91	0.91
set 11	**84.85**	0.86	0.85	0.84	**93.71**	0.94	0.94	0.94
set 12	**84.85**	0.86	0.85	0.83	92.82	0.93	0.93	0.93
set 13	**86.36**	0.89	0.86	0.86	92.31	0.92	0.92	0.92
set 14	**84.85**	0.86	0.85	0.83	92.95	0.93	0.93	0.93
set 15	**88.64**	0.90	0.89	0.88	92.82	0.93	0.93	0.93
set 16	**84.85**	0.87	0.85	0.84	**93.52**	0.94	0.94	0.94
set 17	84.09	0.86	0.84	0.83	**93.46**	0.93	0.93	0.93
set 18	**85.61**	0.88	0.86	0.84	**94.22**	0.94	0.94	0.94
set 19	**86.36**	0.88	0.86	0.85	**94.28**	0.94	0.94	0.94
set 20	**87.12**	0.88	0.87	0.86	92.63	0.93	0.93	0.93
set 21	**85.61**	0.87	0.86	0.84	92.82	0.93	0.93	0.93
set 22	**85.61**	0.87	0.86	0.84	92.88	0.93	0.93	0.93
set 23	**84.85**	0.86	0.85	0.84	**94.09**	0.94	0.94	0.94

TABLE VI

THIS TABLE DESCRIBES THE PERFORMANCE METRICS OF ANN CLASSIFIER ON PF-STAR AND CMU KIDS DATASET WITH VARIOUS FEATURE SET COMBINATIONS FOR PCA VARIANCE RATIO 0.94. THE TABLE PRESENTS KEY PERFORMANCE METRICS, INCLUDING ACCURACY (A) IN PERCENTAGE, PRECISION (P), RECALL (R), AND F1 SCORE (F1).

Feature Set	PF-Star				CMU Kids			
	A	P	R	F1	A	P	R	F1
MFCC (Baseline)	84.09	0.85	0.84	0.83	93.27	0.93	0.93	0.93
set 1	**87.12**	0.90	0.87	0.86	**93.77**	0.94	0.94	0.94
set 2	83.33	0.84	0.83	0.82	**92.69**	0.93	0.93	0.93
set 3	**86.36**	0.87	0.86	0.84	**93.33**	0.93	0.93	0.93
set 4	**85.61**	0.87	0.86	0.84	91.93	0.92	0.92	0.92
set 5	**84.85**	0.86	0.85	0.84	90.47	0.91	0.90	0.90
set 6	**85.61**	0.87	0.86	0.84	**93.58**	0.94	0.94	0.94
set 7	**88.64**	0.90	0.89	0.88	92.25	0.92	0.92	0.92
set 8	**85.61**	0.87	0.86	0.84	92.76	0.93	0.93	0.93
set 9	83.33	0.86	0.83	0.82	92.69	0.93	0.93	0.93
set 10	**86.36**	0.88	0.86	0.85	91.93	0.92	0.92	0.92
set 11	**84.85**	0.87	0.85	0.84	**93.84**	0.94	0.94	0.94
set 12	**84.85**	0.86	0.85	0.83	92.76	0.93	0.93	0.93
set 13	**86.36**	0.89	0.86	0.86	91.93	0.92	0.92	0.92
set 14	**84.85**	0.86	0.85	0.83	92.76	0.93	0.93	0.93
set 15	**88.64**	0.90	0.89	0.88	92.82	0.93	0.93	0.93
set 16	**84.85**	0.87	0.85	0.84	92.88	0.93	0.93	0.93
set 17	**84.85**	0.87	0.85	0.84	93.14	0.93	0.93	0.93
set 18	**85.61**	0.88	0.86	0.84	**93.71**	0.94	0.94	0.94
set 19	**86.36**	0.89	0.86	0.85	**94.79**	0.95	0.95	0.95
set 20	**87.12**	0.88	0.87	0.85	92.82	0.93	0.93	0.93
set 21	**85.61**	0.86	0.86	0.84	92.63	0.93	0.93	0.93
set 22	**85.61**	0.87	0.86	0.84	**93.33**	0.93	0.93	0.93
set 23	**85.61**	0.87	0.86	0.84	**93.9**	0.94	0.94	0.94

VI. CONCLUSION

In this study we have explored various acoustic and prosodic features for children's age classification. The study analyzes 23 feature sets, including the baseline of 26 MFCC coefficients and their combinations with PCA, across PF-Star

and CMU Kids datasets. Notably, for the PF-Star dataset, feature set 1 (MFCC+Chroma), feature set 7 (MFCC+Tempo) and feature set 19 (MFCC+Chroma+Spectral+Pitch+Log Energy), demonstrated the best classification accuracy of 88.64% across three variance ratios (0.98, 0.96, 0.94) during PCA. Whereas for CMU Kids dataset the feature set 19 (MFCC+Chroma+Spectral), achieved the best classification accuracy of 94.73% using an ANN classifier with PCA variance ratio 0.94. The aforementioned unique feature combinations showcased an absolute improvement of 4.55% and 1.46%, in classification of children's age for PF-Star and CMU Kids respectively, over the baseline performances. While currently available state of the art ML and DL classifiers perform well with MFCC features, however, from our analysis, we found that combining additional acoustic and prosodic features to the base MFCC features showed an significant improvement for both PF-Star and CMU Kids test data.

While we already know age classification performance on young children's speech is challenging. Studying the different feature set combinations, we observed that adding acoustics and prosody features to the base MFCC feature proved more beneficial for children age classification task. This comprehensive evaluation provides a nuanced perspective on the effectiveness of the proposed acoustic and prosodic features for age classification, further contributing to the understanding of their applicability in real-world scenarios.

REFERENCES

[1] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech Language*, vol. 27, no. 1, pp. 4–39, 2013, Special issue on Paralinguistics in Naturalistic Speech and Language.

[2] Daria Panek, Andrzej Skalski, Janusz Gajda, and Ryszard Tadeusiewicz, "Acoustic analysis assessment in speech pathology detection," *International Journal of Applied Mathematics and Computer Science*, vol. 25, no. 3, pp. 631–643, 2015.

[3] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[4] Muhammad Ilyas, Alice Othmani, and Amine Naït-Ali, "Auditory perception based system for age classification and estimation using dynamic frequency sound," *Multimedia Tools and Applications*, vol. 79, pp. 21603–21626, 2020.

[5] Sammi Taylor, Christopher Dromey, Shawn L Nissen, Kristine Tanner, Dennis Eggett, and Kim Corbin-Lewis, "Age-related changes in speech and voice: spectral and cepstral measures," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 3, pp. 647–660, 2020.

[6] Davood Mahmoodi, Hossein Marvi, Mehdi Taghizadeh, Ali Soleimani, Farbod Razzazi, and Marzieh Mahmoodi, "Age estimation based on speech features and support vector machine," in *3rd Computer Science and Electronic Engineering Conference (CEEC)*. IEEE, 2011, pp. 60–64.

[7] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[8] Rafik Djemili, Hocine Bourouba, and Mohamed Cherif Amara Korba, "A speech signal based gender identification system using four classifiers," in *International conference on multimedia computing and systems*. IEEE, 2012, pp. 184–187.

[9] Ruben Zazo, Phani Sankar Nidadavolu, Nanxin Chen, Joaquin Gonzalez-Rodriguez, and Najim Dehak, "Age estimation in short speech utterances based on lstm recurrent neural networks," *IEEE Access*, vol. 6, pp. 22524–22530, 2018.

[10] Pegah Ghahremani, Phani Sankar Nidadavolu, Nanxin Chen, Jesús Villalba, Daniel Povey, Sanjeev Khudanpur, and Najim Dehak, "End-to-end deep neural network age estimation.," in *Interspeech*, 2018, pp. 277–281.

[11] Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy, "A deep neural network based end to end model for joint height and age estimation from short duration speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6580–6584.

[12] Hemant K Kathania, Syed Shahnawazuddin, Nagaraj Adiga, and Waquar Ahmad, "Role of prosodic features on children's speech recognition," in *International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5519–5523.

[13] Hemant Kumar Kathania, S Shahnawazuddin, Gayadhar Pradhan, and AB Samaddar, "Experiments on children's speech recognition under acoustically mismatched conditions," in *Region 10 Conference (TENCON)*. IEEE, 2016, pp. 3014–3017.

[14] Martin Russell, "The pf-star british english childrens speech corpus," *The Speech Ark Limited*, 2006.

[15] Maxine Eskenazi, Jack Mostow, and David Graff, "The cmu kids corpus," *Linguistic Data Consortium*, vol. 11, 1997.

[16] Bo Liang, SD Iwnicki, and Yunshi Zhao, "Application of power spectrum, cepstrum, higher order spectrum and neural network analyses for induction motor fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 39, no. 1-2, pp. 342–360, 2013.

[17] Srinivasan Ramakrishnan, "Recognition of emotion from speech: A review," *Speech Enhancement, Modeling and recognition–algorithms and Applications*, vol. 7, pp. 121–137, 2012.

[18] Chadawan Ittichaichareon, Siwat Suksri, and Thaweesak Yingthawornsuk, "Speech recognition using mfcc," in *International conference on computer graphics, simulation and modeling*, 2012, vol. 9.

[19] Dan Ellis, "Chroma feature analysis and synthesis," *Resources of laboratory for the recognition and organization of speech and Audio-LabROSA*, vol. 5, 2007.

[20] Christopher Harte, Mark Sandler, and Martin Gasser, "Detecting harmonic change in musical audio," 2006.

[21] Ahmad R Abu-El-Quran and Rafik A Goubran, "Pitch-based feature extraction for audio classification," in *The 2nd Internatioal Workshop on Haptic, Audio and Visual Environments and Their Applications*. IEEE, 2003, pp. 43–47.

[22] Buket D Barkana and Jingcheng Zhou, "A new pitch-range based feature set for a speaker's age and gender classification," *Applied Acoustics*, vol. 98, pp. 52–61, 2015.

[23] Kuldip K Paliwal, "Spectral subband centroid features for speech recognition," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 1998, vol. 2, pp. 617–620.

[24] Tomi Kinnunen, Bingjun Zhang, Jia Zhu, and Ye Wang, "Speaker verification with adaptive spectral subband centroids," in *Advances in Biometrics*, Seong-Whan Lee and Stan Z. Li, Eds., Berlin, Heidelberg, 2007, pp. 58–66, Springer Berlin Heidelberg.

[25] Jia Min Karen Kua, Tharmarajah Thiruvaran, Mohaddeseh Nosratighods, Eliathamby Ambikairajah, and Julien Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition.," in *Odyssey*, 2010.

[26] Bhargab Medhi and PH Talukdar, "Different acoustic feature parameters zcr, ste, lpc and mfcc analysis of assamese vowel phonemes," in *International Conference on Frontiers in Mathematics*. Gauhati University, 2015, pp. 26–28.

[27] Neha Chauhan, Tsuyoshi Isshiki, and Dongju Li, "Speaker recognition using lpc, mfcc, zcr features with ann and svm classifier for large input database," in *International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2019, pp. 130–133.

[28] Anton Batliner, Andreas Kießling, Ralf Kompe, Heinrich Niemann, and Elmar Nöth, "Tempo and its change in spontaneous speech," 1997.

[29] Guus de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[30] Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky, "Spectral entropy based feature for robust asr," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2004, vol. 1, pp. I–193.