A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 1

Longitudinal study of voice recognition in children

Sandip Purnapatra[1], Priyanka Das[2], Laura Holsopple[3], Stephanie Schuckers[4]

Abstract: Speaker recognition as a biometric modality is on the rise in the consumer marketplace for banking, online services, and personal assistant services with a potential for wider application areas. Most current applications involve adults. One of the biggest challenges in speaker recognition for children is the change in the voice properties as a child age. This work proposes a baseline longitudinal dataset from the same 30 children in the age group of 4 to 14 years over a time frame of 2.5 years and evaluates speaker recognition performance in children with the available speaker recognition technology.

Keywords: Speaker verification, Children's voice, MFCC, LFCC, GMM, JFA, ISV, Inter-session variability.

1 Introduction

Biometric recognition has proliferated in the last two decades with applications in government (border security, immigration, identity at birth, distribution of benefits, refugee efforts) and consumer market (e-commerce, banking, healthcare). Biometric recognition based on voice uses unique features in the speaker's voice to ascertain identity [Ma00]. Voice biometrics uses acoustic properties specific to individual subjects and can be used in situations involving virtual presence over any telephone or internet. Voice biometrics is applied mostly for speaker verification. Speaker verification can be text-dependent or text-independent. For either, the biometric characteristic contains features of the voice specific to a person. Voice biometrics for speaker recognition has been used sparsely since late 1990s. However, in the past decade the application of speaker recognition proliferated in the consumer market for personal assistant services in mobile devices, online services requiring authentication like online banking services, call centers and other services.

Most of the prior research involving voices of children are based on physiological changes of voices with targeted applications like gender recognition, and speech recognition. Speaker recognition performance is still a relatively unexplored research area. One of the few studies in this area shows that as a child ages, their vocal properties changes, impacting the performance of speaker recognition [SRJ18]. The paper is described in more detail at the end of this section.

Studies have shown that developmental speech production, especially vocal tract growth, introduces age-dependent spectral and temporal variability in the speech signal of children. Such variability evoke challenges for robust automatic recognition of children's speech [PN03]. However, no research regarding the influence of vocal tract growth for automatic speaker recognition has been

[1] PhD Student, Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, NY-13676, US, purnaps@clarkson.edu
[2] PhD Student, Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, NY-13676, US, prdas@clarkson.edu
[3] Associate director of CITeR, Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, NY-13676, US, lholsopp@clarkson.edu
[4] Professor, Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, NY-13676, US, sschucke@clarkson.edu

performed for children's voices. According to [Ma00], changes in the voice properties in children add to the basic challenges in voice recognition- background and channel noise; variable and inferior microphones and telephones; and extreme hoarseness, fatigue, or vocal stress.

Extraction of useful information from speech has been researched actively in the last three decades. Mel frequency cepstral coefficient (MFCC), which mimics the frequency response of the human ear, is a well established feature used extensively in most voice/speaker recognition techniques [MBE10]. MFCC filters are designed in accordance to the critical bandwidth frequencies that the human ear perceives. MFCC uses two types of filter- linearly spaced and logarithmically spaced [NS14]. Linear frequency cepstral coefficients (LFCC) is another feature extraction technique that uses only linearly spaced filters. LFCC provides equal details for all frequencies [Re94]. In the higher frequency region of speech, LFCC uses higher number of filterbank compared to MFCC. Inter to intra class speaker variability ratio or f-ratio is significantly higher in LFCC than MFCC [LL09].

In the 1990's many methods like simple template matching, statistical pattern recognition, dynamic time-warping methods were used for speaker recognition. Hidden markov models (HMM), Gaussian mixture model (GMM), universal back ground model (UBM) and multi-layer perceptron gained popularity as speaker recognition techniques in the early 2000's [NS14]. In the last decade, speaker recognition techniques that are based on different types of factor analysis i.e. joint factor analysis (JFA), i-vectors, linear discriminant analysis (LDA) and probabilistic linear discriminant analysis (PLDA) produced improved speaker recognition results [Ka14]. GMM based speaker recognition techniques are not designed to compensate for the inter-session sound variability of different recordings and fails to minimize the the variation in enrollment and probe recordings induced by environmental factors. JFA minimizes these season variability caused by the sound difference of a given speaker's different recordings [Ke07a] [Mc10]. JFA is the Gaussian distribution of HMM supervectors which are speaker and channel dependent and account a few hidden variables of speaker and channel factors or high dimensional GMM super-vectors. JFA model assumes the speaker factor in two different recordings remain same but the channel factor or the recording environment varies from session to session [Ke05] [Ke07a]. JFA models does not work on the speaker verification on short utterance recordings ($<$10 seconds). Rather than modelling the speaker or channel variability space, intermediate-size vector or i-vector models speaker and channel variability in a low dimensional, single total-variability space that can map the utterances (short utterances as well) of the speakers and help convert the speaker recognition problem from a high dimensional to a low dimensional one [Ka14] [De09]. Inter-session variability (ISV) is another modelling approach similar to JFA which aims to reduce inter-session variability in GMM speaker model space [VS08] [Ke07b].The main difference between ISV and JFA is, while ISV modelling assumes that *within subject* variability is dominant in the linear domain of the GMM super vector low-dimensional subspace, JFA assumes that the *between-subject* variability is contained in the low-dimensional subspace [Wa11]. i-vector is also designed to mitigate the speaker variability caused by collecting data from different sources. As i-vectors are computed from the hidden variables of the factor analysis model, it requires huge amount of training data. However, i-vector does not address channel variability; it needs to be combined with other models such as LDA, probabilistic LDA (PLDA), cosine similarity scoring (CSS), within-class covariance normalization (WCCN), which divides the total variability space into session and speaker variability sub-spaces [De09] [Ka11], to mitigate the discrepancy between channel noise of different samples. Deep learning (DL) based speaker recognition systems have the capacity to extract the low-dimensional features and achieve strong speaker recognition performance [Gu20] [Li20b]. Although the DL models produce an improved speaker recognition performance compared to classifiers that require hand-crafted features, they are more complex, requires massive amount of labeled training data, has high computation and storage cost [Li20a].

There has been limited work on speaker verification in children. Safavi et al. [Sa16] [Sa14] [SRJ18] performed automatic speaker, gender and age group identification of approximately 1100 children of different age groups using MFCC, delta and delta-square features and GMM-UBM, GMM-SVM, i-vector-PLDA based models to achieve maximum 99% identification accuracy using only 10 seconds speech recording. However, the dataset details for speaker recognition analysis did not mention

a multi-session collection. More study is needed where the enrollment and verification happen on different days and in multiple sessions spaced by considerable time gap, particularly as a child ages. To the best of our knowledge no report has been published on longitudinal voice biometric recordings for speaker recognition in children. Our study is the first work evaluating speaker verification performance in children with data collected over multi-sessions. In this study, we analyzed longitudinal speaker verification performance in children over a period of 2.5 years with data collected from six sessions with inter-session gap of six months, in time frames of 6, 12, 18, 24 and 30 months between enrollment and verification samples, for the age group of 4 to 14 years using the available technology for adult speaker verification, with approximate recording duration of 90 seconds per subject per session. We report on the longitudinal robustness of speaker verification in children as they age. This work contributes to the research domain by-

1. providing a baseline longitudinal dataset for speaker recognition in children to advance research in this field;
2. evaluating the robustness of established techniques for speaker recognition with child voice data;
3. analyzing the longitudinal speaker verification performance in children.

The rest of the paper is organized in four sections- Section 2 explains data collection protocol, Section 3 details the experimentation steps, Section 4 highlights the results achieved and Section 5 provides a discussion of the limitations of this study, future scopes and concludes on the feasibility of the state-of-art techniques of voice recognition in children.

2 Dataset

The dataset consists of data from the same 30 subjects, for over 2.5 years period, collected from six sessions at an approximated interval of six months from subjects aged between 4 and 11 yrs at enrollment. Using the first session data as the enrollment, longitudinal performance of the dataset has been tested for five subsequent time instances at 6, 12, 18,24 and 30 months. Subject count for each enrollment age between 4 and 11 years are 1, 1, 5, 3, 6, 2, 8 and 4, respectively. The data used for this study is part of a multi-modal biometric dataset collected from the same children for research purpose in cooperation with a local school. The research team sets up collection stations at the school every six months for the collection days using the same equipment. The collection room may vary based on availability which may impact the data.

Voice data is collected at a sampling rate of 44.1 KHz using a microphone by Audio-Technica with frequency response 20Hz to 16KHz and bit depth of 16bit and a publicly available software, Audacity. At each session the subjects are prompted by a series of images to speak simple common words like numbering (1-10), name of animals and common objects known to children and at the end they are asked to describe a scene displayed to them as an image. The speech duration varies based on the speaking speed of the subjects including pauses in between words. Only one sample is collected at each session from each subject of approximate duration of 90 seconds. The protocol and the content was same in all collections. However, the order of images, and thus the words, may have varied. This study focuses on text-independent verification and the content and the order are not considered. Since the data is collected in a school environment, even with our best effort, the collected data have inconsistent noise ranging from sound of people walking, opening or closing of doors, and people talking nearby.

3 Experimentation

3.1 Experimentation Platform: Bob

Bob [An12] [An17] is an open source, reproducible signal processing toolbox. *bob.bio.spear* is a speaker recognition package in the Bob platform having supporting tools for speech data pre-processing, feature extraction, matching and analysis. All experiments with our child voice data has been performed in this platform.

3.2 Data Pre-processing

The data collected for this experiment is in a real life scenario i.e, the data includes channel noise from devices and other environmental noise. Practical applications may not include noise free environment. Thus, it is important to pre-process the data without losing the voice print and distorting the features Our pre-processing includes three steps:

1. Band pass filtering between range 125 Hz and 8000 Hz
2. Mean-Variance Normalization
3. Silence Removal by identifying- (a) Short-time Energy and (b) Spectral Centroid

These data pre-processing steps were performed in MATLAB 2019a, prior to our experimentation in Bob. The Bob experimentation also includes data pre-processing. For our experimentation we used the inbuilt pre-processing resource- *energy_thr*[ID], which is a thresholded energy based voice detection function. The default threshold is 15% of the maximum energy of the input signal, which was used for the secondary pre-processing of our data in the Bob platform. No data was removed due to quality or noise purpose before experimentation in the Bob platform. However, the pre-processor used in the bob-platform failed to process eight samples from eight different subjects at random sessions.

3.3 Feature Extraction and Algorithm

State of art features and algorithms were tested to assess longitudinal speaker recognition performance in children over 2.5 years. Two different feature sets- MFCC and LFCC, were tested with 20 and 60 coefficients for both the feature extraction techniques. Three algorithms- GMM, ISV and JFA, were used to assess performance. Speaker recognition performance from 12 different feature-algorithm combinations tested for our study are tabulated in Table 1.

4 Results and Analysis

Performance is evaluated in terms of False Accept Rate (FAR), False Reject Rate (FRR) and Equal Error Rate (EER). Figure 1 - 12 shows the score distributions and Figure 13 - 24 shows the ROCs for 12 different feature-algorithm combinations for each five longitudinal time instances (6,12,18,24 and 30 months) for the same 30 subjects. Table 1 summarizes the performance at each time instances for each of 12 combination of feature-algorithm in terms of EER.

With MFCC60 and LFCC60, there is decaying variability in the score distribution with ISV algorithm (refer Fig. 4, 10). ISV was reported in literature to have improved speaker verification performance in adults [VS08]. However, we note a drastic degradation in performance with our children

Tab. 1: Speaker Verification performance

Feature	Algorithm	EER (%) 6 month	EER (%) 12 month	EER (%) 18 month	EER (%) 24 month	EER (%) 30 month
MFCC 20	GMM	22	26	30	24	42
MFCC 20	ISV	48	46	56	52	54
MFCC 20	JFA	34	38	35	40	43
MFCC 60	GMM	36	38	40	43	42.5
MFCC 60	ISV	36	44	40	46	46
MFCC 60	JFA	43	37	44	46	52
LFCC 20	GMM	26	34	29	40	48
LFCC 20	ISV	48	47	50	59	56
LFCC 20	JFA	43	38	45	44	50
LFCC 60	GMM	38	35	41	45	51
LFCC 60	ISV	48	52	46	52	54
LFCC 60	JFA	44.5	36	42	52	47.5

dataset as reflected in the ROCs (refer Fig. 15, 16, 21, 22). Though the performance improves for 60 feature dimension compared to 20 feature dimension, the performance of ISV is poor compared to both JFA and GMM (refer 1).

Joint Factor Analysis, which is an extension of ISV, is designed to reduce inter-session variability for intra-subject data and to reduce the high enrollment requirement. The reduced inter-session variability is reflected in the score distributions in Figure 5, 6, 11 as well as in the reduced variability in the performance between longitudinal time instances (6, 12, 18, 24 and 30 months). However, the overall performance is poor compared to GMM with the same set of features.

MFCC20, MFCC60, LFCC20 and LFCC60 features has high variability across longitudinal time instances (6,12,18,24 and 30) with GMM. There is a distinct decay in genuine match scores with GMM for MFCC20 and LFCC20 (refer Fig. 1, 7). The score distribution for 60 dimensional features for MFCC and LFCC show higher variability. However, GMM performs best with all 4 configurations- MFCC-20, MFCC60, LFCC-20, LFCC-60, compared to ISV and JFA. The best performance is observed for the MFCC20 and GMM combination in terms of FAR and FRR (EER varies from 22% at 6 month time instance to 42% at 30 month time instance) compared to other algorithms and features. Overall, 20 dimensional feature vector for both MFCC and LFCC perform better compared to 60 dimensional features. Almost all feature-algorithm combination fails to perform at 30 month time frame with EER ranging between 42% to 56 %.

5 Discussion, Limitation and Future Scope

In the last few years several speaker verification systems has been proposed. However, impact of increased time between voice enrollment and probe samples on speaker recognition performance are still an unexplored area, especially in children. This work is an attempt to answer the question- *Are the available voice recognition techniques robust enough to recognize children as they age?* For this purpose, a dataset has been collected from the same 30 children in six sessions over 2.5 years. The data has been analyzed using state of art features and algorithms that have proved effective for adult speaker verification.

From our analysis, we conclude that MFCC20 features and GMM algorithm performs best for longitudinal speaker verification in children. However, the best performance is not on par with the

Score distribution for genuine matches as time increases between enrollment and verification

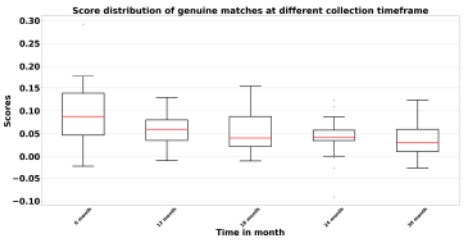

Fig. 1: Feature: MFCC20; Algo: GMM

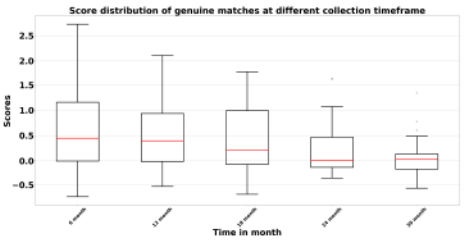

Fig. 2: Feature: MFCC60; Algo: GMM

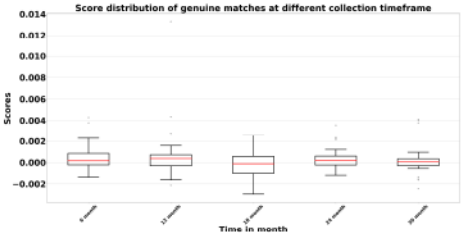

Fig. 3: Feature: MFCC20; Algo: ISV

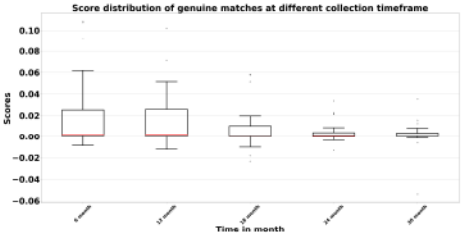

Fig. 4: Feature: MFCC60; Algo: ISV

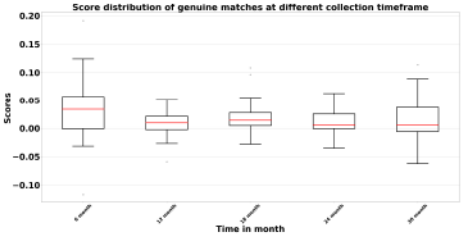

Fig. 5: Feature: MFCC20; Algo: JFA

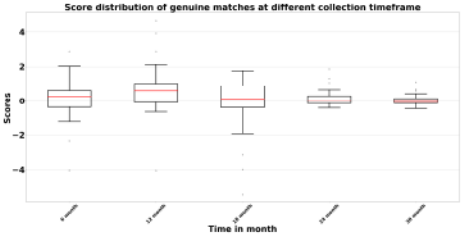

Fig. 6: Feature: MFCC60; Algo: JFA

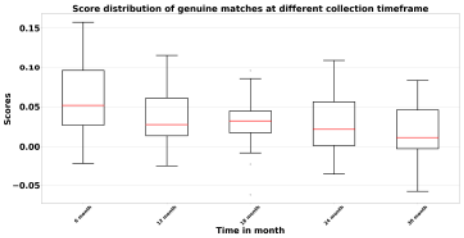

Fig. 7: Feature: LFCC20; Algo: GMM

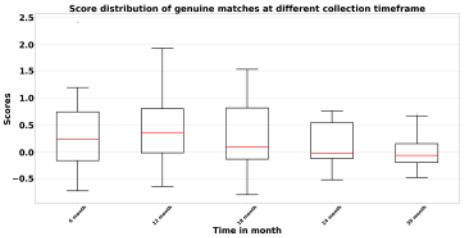

Fig. 8: Feature: LFCC60; Algo: GMM

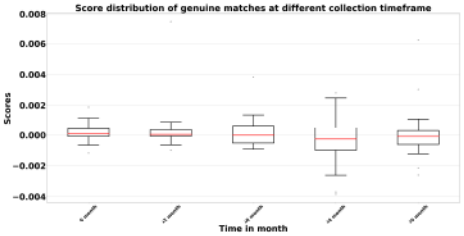

Fig. 9: Feature: LFCC20; Algo: ISV

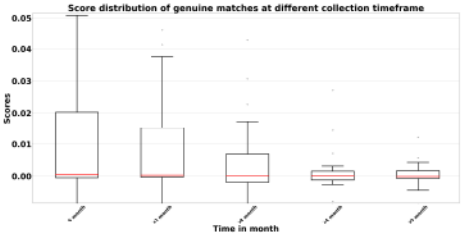

Fig. 10: Feature: LFCC60; Algo: ISV

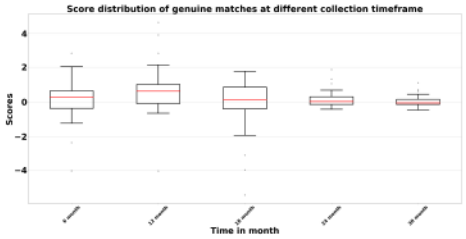

Fig. 11: Feature: LFCC20; Algo: JFA

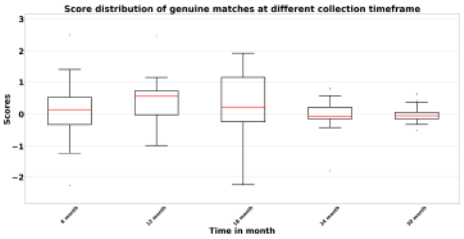

Fig. 12: Feature: LFCC60; Algo: JFA

expected biometric recognition performance. The state of art algorithms (ISV and JFA) designed to reduce inter session variability and improve recognition performance, do perform well in children. However, these are not commercially developed algorithms, which we assume might perform differently. We note that there is need for improvement of speaker recognition in children with the development of appropriate features and algorithms.

The data used in the study was collected in a real life scenario with background noise including sound of people walking by, talking, opening and closing of doors and other miscellaneous noise. However, not all data for all subjects have noise and the noise level varies between sessions and subjects. No complete session was deleted due to noise. We removed pauses between utterances to reduce noise in the data. Most noise frequencies are in the range of human voice frequencies. Thus, even with the best effort it was not possible to eliminate noise frequencies from the signal without effecting the voice properties. Thus it is expected to have degraded performance in recognition compared to ideal voice samples. To the best of our knowledge, no publicly available multi-session voice dataset from children is available to support research in this field. Pre-trained networks on adult data has proved inefficient when used in applications involving children for other modalities like face, where high variability is observed with aging [DNJ18]. However, it can be a work for future to test the viability of such approach with child speaker verification. State of art algorithms with hand crafted features do not require a large amount of data for supervised training has proved high efficiency. Cases with limited amount of data needs robust algorithm pipeline for applications in terms of both features and algorithm. We recognize that non-availability of dataset is a hindrance to our research community. We also recognize privacy and sensitivity related to child biometric data. We are in the process of sharing our dataset through BEAT platform to support research in this field while protecting data privacy. All algorithms used for analysis are also available through an interface from BEAT to the Bob platform.

This work initiates research in the field of child voice recognition impacted by aging. For future work statistical modelling of the variation in voice signature features may help in modelling biometric aging in child voices. The very basis of biometrics is temporal-stability. Time-invariant voice features need to be defined for child in order to be useful for biometric applications. Research on robust feature and classification techniques are required to address speaker recognition with intra-class variability due to aging in children. Further research in this field is needed to support widespread application of voice biometrics across all age groups. We conclude that the state of art algorithms for speaker recognition performance in adults does not reflect similarly in the case of speaker recognition in children for the age group of 4 to 14 years. There is a need for development of age-independent features and algorithms for child speaker recognition for longitudinal biometric applications.

6 Acknowledgement

This research was funded by Center for Identification Technology and Research (CITeR) and National Science Foundation (NSF)(Grant # :1650503). The database creation was made possible by

ROCs as time increases between enrollment and verification with different feature sets and algorithms

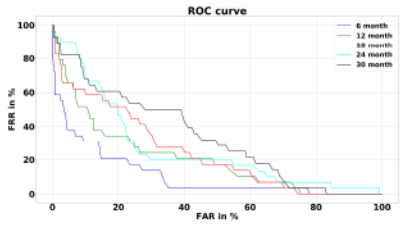

Fig. 13: Feature: MFCC20; Algo: GMM

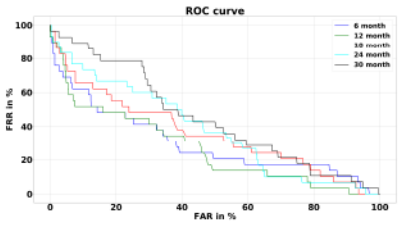

Fig. 14: Feature: MFCC60; Algo: GMM

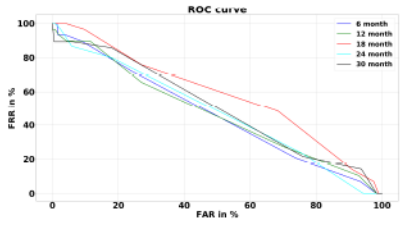

Fig. 15: Feature: MFCC20; Algo: ISV

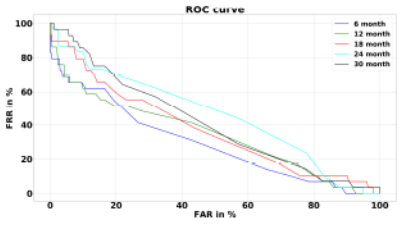

Fig. 16: Feature: MFCC60; Algo: ISV

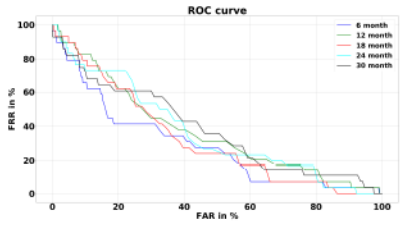

Fig. 17: Feature: MFCC20; Algo: JFA

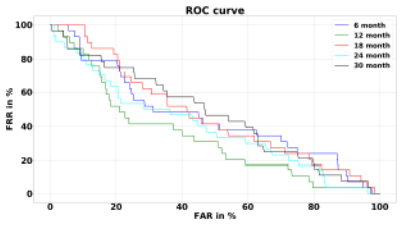

Fig. 18: Feature: MFCC60; Algo: JFA

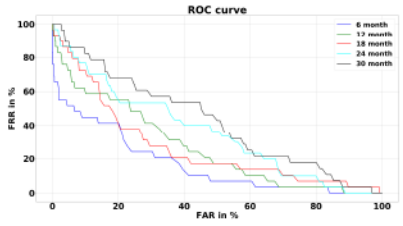

Fig. 19: Feature: LFCC20; Algo: GMM

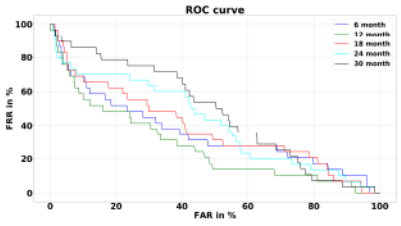

Fig. 20: Feature: LFCC60; Algo: GMM

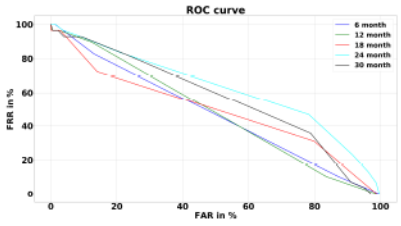

Fig. 21: Feature: LFCC20; Algo: ISV

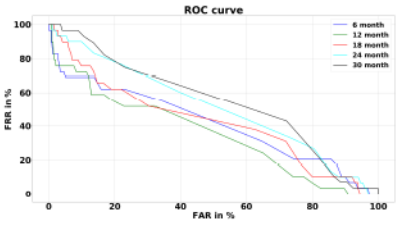

Fig. 22: Feature: LFCC60; Algo: ISV

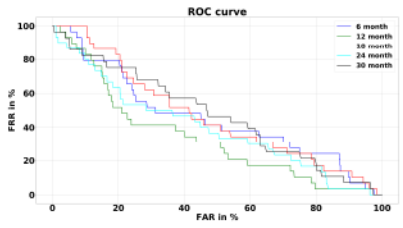

Fig. 23: Feature: LFCC20; Algo: JFA

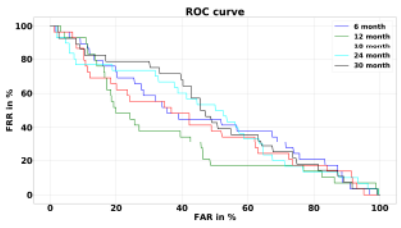

Fig. 24: Feature: LFCC60; Algo: JFA

voluntary participation of all enrolled subjects in the study, their parents/guardians and the hard work of the data collecting team from Clarkson University. We would also extend our gratitude to the Potsdam Elementary School and Potsdam Middle School administration and staff members for their continued support in academic research.

References

[An12] Anjos, A.; Shafey, L. El; Wallace, R.; Günther, M.; McCool, C.; Marcel, S.: Bob: a free signal processing and machine learning toolbox for researchers. In: 20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan. October 2012.

[An17] Anjos, A.; Günther, M.; de Freitas Pereira, T.; Korshunov, P.; Mohammadi, A.; Marcel, S.: Continuously Reproducing Toolchains in Pattern Recognition and Machine Learning Experiments. In: International Conference on Machine Learning (ICML). August 2017.

[De09] Dehak, Najim; Dehak, Reda; Kenny, Patrick; Brümmer, Niko; Ouellet, Pierre; Dumouchel, Pierre: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Tenth Annual conference of the international speech communication association. 2009.

[DNJ18] Deb, Debayan; Nain, Neeta; Jain, Anil K: Longitudinal study of child face recognition. In: 2018 International Conference on Biometrics (ICB). IEEE, pp. 225–232, 2018.

[Gu20] Gusev, Aleksei; Volokhov, Vladimir; Andzhukaev, Tseren; Novoselov, Sergey; Lavrentyeva, Galina; Volkova, Marina; Gazizullina, Alice; Shulipa, Andrey; Gorlanov, Artem; Avdeeva, Anastasia et al.: Deep speaker embeddings for far-field speaker recognition on short utterances. arXiv preprint arXiv:2002.06033, 2020.

[ID] IDIAP: , Energy Theshold Pre-processor. https://pydoc.net/bob.bio.spear/3.1.0/bob.bio.spear.preprocessor.Energy_Thr/. Accessed: 2020-08-17.

[Ka11] Kanagasundaram, Ahilan; Vogt, Robbie; Dean, David B; Sridharan, Sridha; Mason, Michael W: I-vector based speaker recognition on short utterances. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association. International Speech Communication Association (ISCA), pp. 2341–2344, 2011.

[Ka14] Kanagasundaram, Ahilan; Dean, David; Sridharan, Sridha; Gonzalez-Dominguez, Javier; Gonzalez-Rodriguez, Joaquín; Ramos, Daniel: Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. Speech Communication, 59:69–82, 2014.

[Ke05] Kenny, Patrick: Joint factor analysis of speaker and session variability: Theory and algorithms. CRIM, Montreal,(Report) CRIM-06/08-13, 14:28–29, 2005.

[Ke07a] Kenny, Patrick; Boulianne, Gilles; Ouellet, Pierre; Dumouchel, Pierre: Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing, 15(4):1435–1447, 2007.

[Ke07b] Kenny, Patrick; Boulianne, Gilles; Ouellet, Pierre; Dumouchel, Pierre: Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 15(4):1448–1460, 2007.

[Li20a] Li, Ruirui; Jiang, Jyun-Yu; Li, Jiahao Liu; Hsieh, Chu-Cheng; Wang, Wei: Automatic speaker recognition with limited data. In: Proceedings of the 13th International Conference on Web Search and Data Mining. pp. 340–348, 2020.

[Li20b] Liang, Tianyu; Liu, Yi; Xu, Can; Zhang, Xianwei; He, Liang: Combined Vector Based on Factorized Time-delay Neural Network for Text-Independent Speaker Recognition. In: Proc. Odyssey 2020 The Speaker and Language Recognition Workshop. pp. 428–432, 2020.

[LL09] Lei, Howard; Lopez, Eduardo: Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In: Tenth Annual Conference of the International Speech Communication Association. 2009.

[Ma00] Markowitz, Judith A: Voice biometrics. Communications of the ACM, 43(9):66–73, 2000.

[MBE10] Muda, Lindasalwa; Begam, Mumtaj; Elamvazuthi, Irraivan: Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083, 2010.

[Mc10] McLaren, Mitchell; Vogt, Robert; Baker, Brendan; Sridharan, Sridha: A comparison of session variability compensation approaches for speaker verification. IEEE Transactions on Information Forensics and Security, 5(4):802–809, 2010.

[NS14] Nijhawan, Geeta; Soni, MK: Speaker recognition using support vector machine. International Journal of Computer Applications, 87(2), 2014.

[PN03] Potamianos, Alexandros; Narayanan, Shrikanth: Robust recognition of children's speech. IEEE Transactions on speech and audio processing, 11(6):603–616, 2003.

[Re94] Reynolds, Douglas A: Experimental evaluation of features for robust speaker identification. IEEE Transactions on Speech and Audio Processing, 2(4):639–643, 1994.

[Sa14] Safavi, Saeid; Najafian, Maryam; Hanani, Abualsoud; Russell, Martin J; Jancovic, Peter: Comparison of speaker verification performance for adult and child speech. In: WOCCI. pp. 27–31, 2014.

[Sa16] Safavi, Saeid; Najafian, Maryam; Hanani, Abualsoud; Russell, Martin J; Jancovic, Peter; Carey, Michael J: Speaker recognition for children's speech. arXiv preprint arXiv:1609.07498, 2016.

[SRJ18] Safavi, Saeid; Russell, Martin; Jančovič, Peter: Automatic speaker, age-group and gender identification from children's speech. Computer Speech & Language, 50:141–156, 2018.

[VS08] Vogt, Robbie; Sridharan, Sridha: Explicit modelling of session variability for speaker verification. Computer Speech & Language, 22(1):17–38, 2008.

[Wa11] Wallace, Roy; McLaren, Mitchell; McCool, Christopher; Marcel, Sebastien: Inter-session variability modelling and joint factor analysis for face authentication. In: 2011 International Joint Conference on Biometrics (IJCB). IEEE, pp. 1–8, 2011.