Interspeech 2018
2-6 September 2018, Hyderabad

Fusing text-dependent word-level i-Vector models to screen 'at risk' child speech

Prasanna Kothalkar[1], Johanna Rudolph[2], Christine Dollaghan[2], Jennifer McGlothlin[2], Thomas Campbell[2], John H. L. Hansen[1]

[1]Center for Robust Speech Systems, University of Texas at Dallas
[2]Callier Center for Communication Disorders, University of Texas at Dallas

{prasanna.kothalkar, johanna.rudolph, dollaghan, jmcglothlin, thomas.f.campbell, john.hansen} @utdallas.edu

Abstract

Speech sound disorders (SSDs) are the most prevalent type of communication disorder among preschoolers. The earlier an SSD is identified, the earlier an intervention can be provided to potentially reduce the social/academic impact of the disorder. The challenge, lies in early identification of such disorders. In this study 29 carefully selected words were produced by 165 children from 3-6 years of age. The audio recordings, were collected by parents using a mobile application /platform. "Ground truth" child status as 'typically developing' vs 'at risk' was based on a percentage of consonants correct-revised growth curve model. State-of-the-art speech processing/speaker recognition models were employed along with our clinical group verification framework. Results showed that text-dependent i-Vector models were superior to both text dependent and text-independent Gaussian Mixture Models (GMMs) for correct classification of children. Fusing individual word, i-Vector models provides insight into word and consonant groupings that are more indicative of 'at risk' child speech.

Index Terms: speech processing, machine learning, clinical screening, automatic screening, child speech, speech disorders, voice disorders, speaker verification, growth curve model, mobile application

1. Introduction

Recent advancements in Speech[1, 2] and Speaker Recognition [3] have brought speech technology within the realm of human performance. Voice based interfaces to portable devices like Apple's Siri, Google Voice Recognition, Amazon's Alexa and Microsoft's Cortana have introduced a new age of hands-free human-computer interaction. Speech Recognition involves recognizing sounds produced at the phoneme, word or sentence level. Speaker recognition can include speaker identification or speaker verification. In speaker identification, we determine if a given speaker belongs to a known (closed-set) or unknown (open-set) group of speakers, while in speaker verification, we test if a given speech utterance belongs to the specified speaker. For speaker verification we need two models, the speaker model and an impostor model. The impostor model is comprised of speakers other than the test speaker and is also known as the Universal Background Model (UBM). The model with the higher score is considered to be the test speaker's model. The UBM can be seen as the background data for initializing base parameters and has also been used for speaker dependent adaptation of speaker and impostor models.

Apart from voice search and dialing, biometrics and law enforcement, these technologies are used in healthcare for medical dictation/transcription, speech synthesis for people with impaired speech, text-to-speech for those who are visually impaired, understanding subject or pathological traits through speech etc. Speech processing and machine learning techniques have been used to detect speech disorders and/or their co-morbid conditions such as Alzheimers disease [4], Parkinsons disease [5, 6, 7], Amyotrophic Lateral Sclerosis (ALS) [8], mild-Traumatic Brain Injury (m-TBI) or Concussion [9]. Most of these techniques involve extracting speech features and applying some machine learning models to them in order to detect pathological speech and thus to identify the disorder or medical condition. Measures such as Speech Intelligibility Rate [6, 10] (speech intelligibility X speaking rate) and Unified Parkinson Disease Rating Scale score [7] (questionnaire based score) have also been predicted using Support Vector Regression and Deep Neural Network Regression. Similarly, isolated vowel sounds have been used to predict the presence of concussion [9] using Support Vector Machine. These technologies have also resulted in development of portable devices [5] to process signals and provide feedback for speech therapy. Such innovations in healthcare utilization, coupled with the success of speech and speaker recognition technology as mentioned above, hold great promise for applying these mature systems and algorithms to problems in the clinical domain.

Speech Sound Disorders (SSDs), in particular, affect between 3-16% of US children [11] and could be due to a variety of causes. SSDs may be associated with dysarthria, apraxia, cleft palate, down syndrome, or autism. However, the majority of SSDs in children are idiopathic, that is, there is no known cause or co-morbidity. Most young children exhibit speech sound errors that are developmental and will resolve spontaneously over time. Differentiating between developmental errors that will resolve and errors that indicate an SSD requires considerable training and practice. Thus, there is a need for a portable, accurate, well-researched and simple screening tool to ensure that children with SSDs are identified as early as possible. Population-based 'clinical group' modeling knowledge will help clinicians manage and monitor children who are 'at risk' for SSDs.

The goal of this work was to determine if state-of-the-art speaker recognition feature representations can provide significant improvements over GMMs[12] in objectively identifying children who warrant further monitoring or in-depth clinical evaluation. We utilize the text-dependent setting for generating i-Vectors as it has resolved content mismatch [13, 14] issues between enrollment and test data for short utterances and performs well at word [15] level. To the best of our knowledge, this work represents the first attempt at screening children who

10.21437/Interspeech.2018-1465

are 'at risk' of SSDs using text-dependent i-Vectors following a speaker verification framework. Additionally, we fuse groups of words with similar characteristics to create a more representative test for our system, which can also identify susceptible phoneme classes among children with SSDs.

2. Child Speech Dataset

2.1. Data Collection

We analyzed audio recordings of 29 words from 165 children, between 3-6 years of age, collected by their parents. All participants were living in the Dallas metropolitan area. The Institutional Review Board at the University of Texas at Dallas Office of Research Compliance has granted approval for enrollment of up to 500 participants for this project. The recordings were collected in MPEG-4 AAC format at 44 kHz sampling rate with 96 kbps bit rate using an iOS application developed by our team. Figure 1 shows an example of the interface used to collect recordings within the application. Visual and orthographic representations of the words inform the users of the targets. When the red 'record' button is pushed, an auditory prompt asks: 'What do you call this?' followed by a double beep, which identifies the beginning of the recording. The completed recordings are automatically transferred to a secure Amazon server (Amazon Simple Storage Service).

Measure	Mean	SD	Range
Age (months)	51	11	36-78
PCC (%)	83	16	10-100

Table 1: *Descriptive statistics of the sample*

2.2. Speech Science behind Words and Ground Truth Labeling

Target productions included 14 consonant-vowel-consonant (CVC) monosyllabic words (e.g., hat, soap), four /s/ cluster monosyllabic words (e.g., spoon, star), and 11 multisyllabic words (e.g., caterpillar, elephant) to challenge the child's developing vocabulary. These words include a combination of early, mid, and late developing consonants and key consonant clusters, which have been shown to discriminate between children with and without SSDs. A certified speech-language pathologist transcribed each participants word productions and calculated the percentage of consonants correct (PCC)[16]. A growth curve model [17] was used to determine age-based cutoff scores for PCC. Children whose scores fell more than 1.5 standard deviations below the mean for their age were classified as 'at-risk' or 'SSD'; this included 64 out of the 165 children (39% of the sample).

2.3. Data Preprocessing

Each child provided 20-29 recordings each being 1-10 sec long. Noisy utterances with parents talking, other children in the background, toys etc. or use of articles/phrases different from the target word or repeated words were manually clipped to the recording of the target word. Also incorrect words and files with no sound, were discarded. Our final dataset had 4685 total words from 165 children.

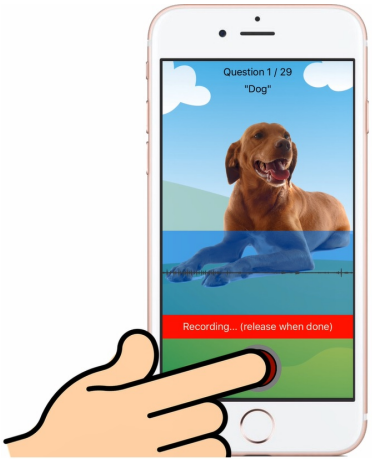

Figure 1: *Illustrative snapshot of mobile application while recording child speech.*

3. Method

Once data pre-processing is complete, Mel-Frequency Cepstral Coefficient (MFCC) features are extracted from the raw audio. Then, the features are transformed to an i-Vector representation. The i-Vectors are used to classify the utterance as belonging to the SSD category or the normal speech acquisition (NSA) category using L2-Logistic Regression and Gaussian Backend classifiers. For word-level modeling, individual utterance scores are compared through a group verification framework. Finally, every word is grouped into different word and consonant categories. Scores of the words belonging to the same category are fused within the same group verification framework and this process provides further results. Our toolkit Multi Session-Acoustic Identification (MS-AcID)[18] is used for each of the following steps.

3.1. Feature Extraction and i-Vector representation

Thirteen dimensional MFCC including delta and double delta features are used. A GMM with 256 mixture components is learned using the MFCC features which is further transformed to i-Vector representation of 50 dimensions. i-Vectors represent speech GMM parameters by decomposing them in terms of speaker dependent and speaker independent components. The speaker independent component is the GMM-Universal Background Model (UBM, represented by m) and consists of a sample of data that includes a different set of users than those used for development and testing. The speaker dependent component includes the 'total variabilty extractor' matrix (T) which extracts identity-Vector or i-Vector (w) which is unique for each speaker. Thus, every speech utterance representation (M) can be expressed as combination of both components.

$$M = m + Tw \tag{1}$$

3.2. Backend Models

3.2.1. L2-Regualrized Logistic Regression (L2LR)

Previous investigations [18] on L2LR for multi-session speaker verification scenario using MFCC features showed competitive results. Given a pair of i-Vector features and clinical group c,

we follow a log-linear probability model to learn weights v in order to classify i-Vectors w into NSA (-1) or SSD (1) class.

$$P(c = \pm 1|w, v) = \frac{1}{1 + \exp(-c(v^T w))} \quad (2)$$

3.2.2. Gaussian Backend (GB)

Gaussian Backend classifier models the i-Vectors for both the clinical groups as a Gaussian distribution and measures the posterior log-likelihood for each group c as:

$$\log p(c|w) = -\frac{1}{2} w^T \Sigma^{-1} w + w^T \Sigma^{-1} m_c - \frac{1}{2} m_c^T \Sigma^{-1} m_c + k \quad (3)$$

where w are the i-Vectors, Σ is the shared covariance matrix for both the clinical groups, m_c is the mean matrix for clinical group c and k is a constant.

3.3. Clinical Group Verification

Our clinical group verification framework[12] extends speaker verification to the group level by transforming utterance level scores to clinical group level. For this, we sum individual word scores for the SSD and NSA models and assign the utterance to the model with the higher score. This transformation also involves converting the likelihood value scores to distance measures which is the absolute difference between the min weighted error rate threshold[19] (MWERTH) and our score. MWERTH measured using BOB toolkit[20], is calculated as follows,

$$MWERTH = cost \times FAR + (1 - cost) \times FRR \quad (4)$$

where FAR stands for False Acceptance Rate, FRR stands for False Rejection Rate, cost is parameter we assign in range (0.4,0.6) with step 0.1.

3.4. Data splits for round robin

Each of the 165 children are randomly split into sets of 32, 32, 33, 33, 35 and each of the 29 words provide 29 such groupings. If a child is missing a word in it's audio collection due to missing or incorrect recording, that child cannot be included for modeling using that word. We train five different models with $(3/5)^{th}$ data in training, and ensure each set is used for validation and testing exactly one time.

4. Results and Discussion

Results over each of the five test sets inferred after cross-validation are presented in terms of sensitivity[21] (true positive rate), specificity[21] (true negative rate) and accuracy. Sensitivity represents the fraction of children who are 'at risk' and are correctly predicted as such. Specificity represents the fraction of children who are 'typically developing' and are predicted as such. Accuracy stands for the total fraction of children that have been correctly predicted based on PCC "ground truth" classification.

4.1. Individual word

Figure 2 reports the maximum accuracy obtained for an individual word in classifying the children, among different combinations of cost and classifier. Words 'hat' (96.83%), 'juice', 'cat', 'elephant' and 'nose' provide the best results in terms of sensitivity and all of these words except 'juice' (6^{th}) are in top 5 in terms of accuracy as well.

4.2. Consonant and word grouping categories

We calculate scores for different combinations of words containing early developing, middle developing, late developing consonants (164 children), s-clusters (163 children), multisyllabic words, high accuracy group and low accuracy group words by adding the predictions of individual words. This resolves a lot of the tie predictions and we are able to compare metric values for different groups to understand their effectiveness in discerning 'typically developing' vs 'at risk' child speech. As can be seen in Figure 3, sensitivity was highest for words with 's-clusters' and words with later-developing consonants; sensitivity for the remaining word types was substantially lower. These groupings provide more balanced results in the form of considering words which might have been discarded due to ties for more than 25% utterances.

4.3. Explaining the ties

The Gaussian Mixture Model which is used to create i-Vectors for single word phoneme-diversity data, uses a large number of mixtures (256) in order to represent every child's varied representation. However, for some words this causes sparsity of the 'i-Vector/ total variability extractor' matrix, not generating i-Vectors for the particular test set cross validation. Words with a 'variability matrix sparsity problem' for less than 75% of the utterance predictions of that word are considered for the model comparison. Utterances of words within this group without i-Vectors, default to being classified as 'Ties' by our verification framework. With the goal of ensuring lower false negatives, we would classify these children as 'at risk'. Thus, we have also provided the tie accuracy (Figures 2, 3) of such classifications and these are around 39% (except for Hat in figure 2).

4.4. Comparison with prior work

Our previous work[12] on screening 'at risk' child speech using text-independent Gaussian Mixture Models (1024 mixtures, 60 dimensional MFCC features) peaked at an accuracy level of 79.88%, sensitivity level of 68.25% and specificity level of 87.13%. These results were better than text-independent i-Vector models. All the word-level text-dependent i-Vector models (as shown in Figure 2) perform better in terms of sensitivity and all but Caterpillar and Scissors are better in terms of accuracy. An equivalent comparison of our i-Vector models with text-independent GMMs, in terms of number of mixtures (256) and MFCC dimensions (39) (accuracy = 66.67%, sensitivity=51.56%, specificity=76.24%), shows a better margin of performance for i-Vectors. The best performing text-dependent GMM model (256 mixtures, 39-dimensional MFCC), which is for the word Van (Accuracy=75.32%, Sensitivity=70.49%, Specificity=78.35%), is well below the top performing word-dependent i-Vector models and is consistent with previous work, signifying the benefit of i-Vectors over GMMs in this setting.

5. Conclusions and Future work

The strong speaker footprint provided by word-level i-Vectors on matched content provided good accuracy, sensitivity results for deviations from typical speech, confirming the efficacy of text-dependent representations for dealing with the challenge of short utterances. Additionally, our solution presented the first attempt at using state-of-the-art i-Vector representations to discriminate between developmental speech sound errors and

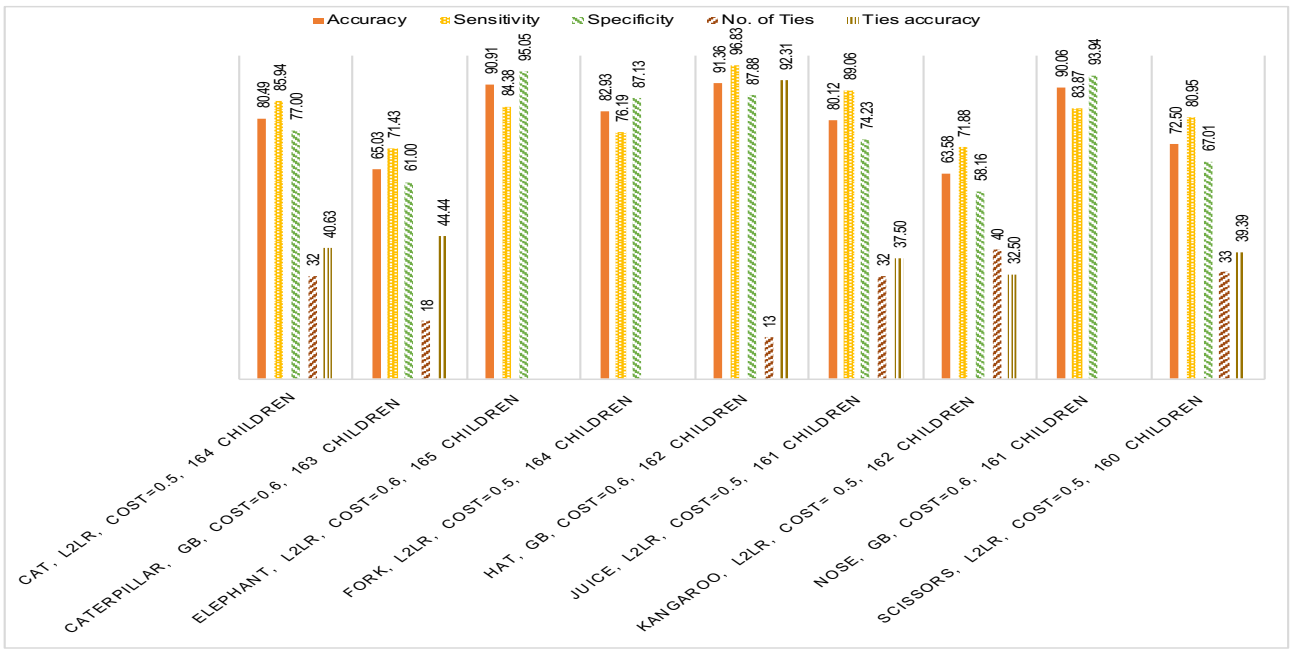

Figure 2: Subject-specific accuracy, sensitivity and specificity on test set for best combination of (classifier, cost in MWERTH) for words Cat, Caterpillar, Elephant, Fork, Hat, Juice, Kangaroo, Nose and Scissors.

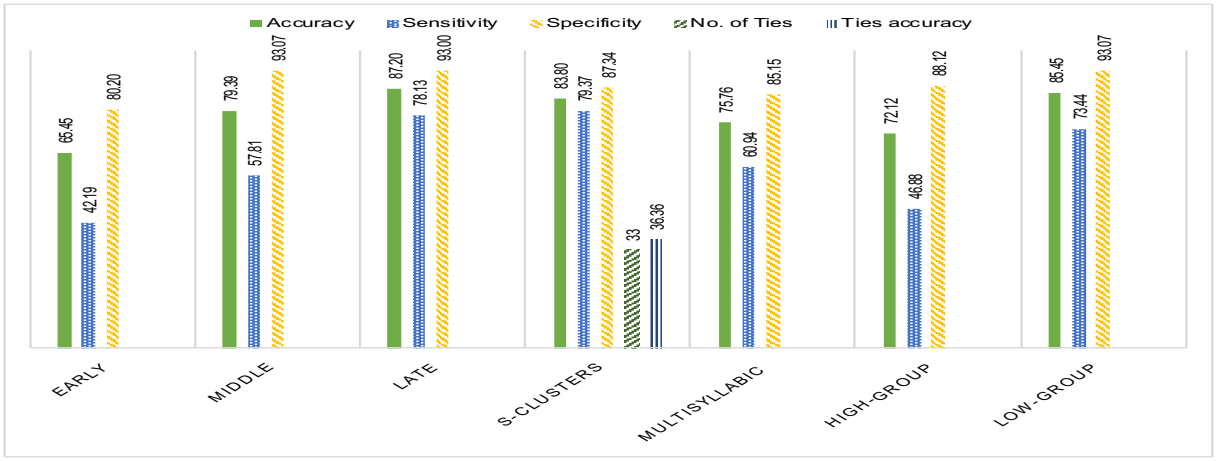

Figure 3: Subject-specific accuracy, sensitivity and specificity on test set for consonant and word grouping categories with cost = 0.5 (best on development set) in MWERTH using L2LR.

errors indicative of a possible speech-sound disorder. Our clinical group verification framework easily adapted to accumulating scores for each word and came up with predictions for each child based on different word groupings. The metrics calculated provided insight for some words and specific phoneme types that were more discerning of 'at risk' child speech. This knowledge could ultimately be useful clinically in monitoring child speech development. It could also be relevant to designing datasets for speech-based artificial intelligence solutions to assist clinicians. We would like to make our research practically viable, by handling noise, automatic parsing etc. From a speech science perspective, we plan further analyses to identify additional speech patterns representative of 'at risk' speech.

6. Acknowledgements

This work was funded by the Callier Center for Communication Disorders Excellence in Education Endowment and the Sara T. Martineau Professorship in Communication Disorders. We would like to thank Dr. Abhijeet Sangwan, Shawn Spencer and the volunteering participants.

7. References

[1] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines," Interspeech 2017, 2017.

[2] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward Human Parity in Conversational Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2410–2423, 2017.

[3] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," in IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 74-99, Nov. 2015. doi: 10.1109/MSP.2015.2462851

[4] K. C. Fraser, F. Rudzicz, and G. Hirst, "Detecting late-life depression in Alzheimer's disease through analysis of speech and language," Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, 2016.

[5] J. C. V. Correa, J. R. O. Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, & E. Nöth "New computer aided device for real time analysis of speech of people with Parkinsons disease," Revista Facultad de Ingenieria Universidad de Antioquia, 87(72), 87-103, 2014

[6] A. Tsanas, M.A. Little, P.E. McSharry and L.O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," Journal of the Royal Society, Interface / the Royal Society, vol. 8, pp. 842-855, Jun 6,. 2011.

[7] S. Hahm & J. Wang, "Parkinson's Condition Estimation using Speech Acoustic and Inversely Mapped Articulatory Data," Interspeech, 513-517, 2015.

[8] J. Wang, P. V. Kothalkar, B. Cao, and D. Heitzman, "Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples," Interspeech 2016, Aug. 2016.

[9] M. Falcone, N. Yadav, C. Poellabauer, and P. Flynn, "Using isolated vowel sounds for classification of Mild Traumatic Brain Injury," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[10] J. Wang, P. V. Kothalkar, M. Kim, Y. Yunusova, T. F. Campbell, D. Heitzman, and J. R. Green, "Predicting Intelligible Speaking Rate in Individuals with Amyotrophic Lateral Sclerosis from a Small Number of Speech Acoustic and Articulatory Samples," SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies, 2016.

[11] National Academies of Sciences, Engineering, and Medicine. 2016. Speech and language disorders in children: Implications for the Social Security Administrations Supplemental Security Income Program. Washington, DC: The National Academies Press. doi: 10.17226/21872

[12] P. Kothalkar, J. Rudolph, C. Dollaghan, J. McGlothlin, T. Campbell and J.H.L. Hansen, "Automatic Screening to Detect 'At Risk' Child Speech Samples Using a Clinical Group Verification Framework," IEEE-EMBC 2018, Honolulu, HI, July 17-21, 2018.

[13] N. Scheffer, Y. Lei, "Content matching for short duration speaker recognition," Interspeech 2014, pp. 1317–1321, Sep. 2014.

[14] S. Dey, S. Madikeri, P. Motlicek, and M. Ferras, "Content Normalization for Text-Dependent Speaker Verification," Interspeech 2017, 2017.

[15] R. Huang and J. Hansen, "Dialect/Accent Classification via Boosted Word Modeling," Proceedings. (ICASSP 05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.

[16] L. D. Shriberg, D. Austin, B. A. Lewis, J. L. Mcsweeny, and D. L. Wilson, "The Percentage of Consonants Correct (PCC) Metric," Journal of Speech Language and Hearing Research, vol. 40, no. 4, p. 708, Jan. 1997.

[17] T. F. Campbell, C. Dollaghan, J. E. Janosky, and P. D. Adelson, "A Performance Curve for Assessing Change in Percentage of Consonants Correct–Revised (PCC-R)," Journal of Speech Language and Hearing Research, vol. 50, no. 4, p. 1110, Jan. 2007.

[18] G. Liu and J. H. L. Hansen, "An Investigation into Back-end Advancements for Speaker Recognition in Multi-Session and Noisy Enrollment Scenarios," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1978–1992, 2014.

[19] S. Bengio, M. Keller & J. Mariéthoz, "The Expected Performance Curve," International Conference on Machine Learning ICML Workshop on ROC Analysis in Machine Learning, 136(1), 1963–1966, 2004.

[20] A. C. A. Anjos, L. El-Shafey, R. Wallace, M. Gnther, C. Mccool, & S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," Proceedings of the 20th ACM international conference on Multimedia - MM 12, 2012.

[21] D. G. Altman & J. M. Bland, "Statistics Notes: Diagnostic tests 1: sensitivity and specificity," Bmj, vol. 308, no. 6943, pp. 1552–1552, Nov. 1994.