6th Workshop on Child Computer Interaction (WOCCI 2017)
13 November 2017, Glasgow, Scotland, UK

Comparison of two scoring method within i-vector framework for speaker recognition from children's speech

Saeid Safavi[1,2]*, and Lily Meng*[1]

[1]School of Engineering and Technology, University of Hertfordshire, UK
[2]School of Electronic, Electrical and Systems Engineering, University of Birmingham, UK
s.safavi@herts.ac.uk

Abstract

Speaker recognition is a well established area for research but it mainly focuses on adult speech. Recent work on children's speech shows that not all the findings from speaker recognition on adult speech are directly applicable on children's speech. There are a variety of applications for speaker recognition from children's speech, for example it could be used as a safeguard for a child during her/his interactions on social media networking websites. It could also be used as one of the main blocks in automatic tutor systems for educational purposes at schools. In this research we have evaluated two scoring method for speaker recognition within the i-vector framework using two simulated environments; in a classroom (contains 30 students) and in a school (contains 288 students). The first method is based on the PLDA scoring approach and the second method is based on the cosine similarity measure. Results show that the first method outperforms the second approach in a simulated school, but it is the other way around for the recognition of a child in a classroom in which the second scoring method performs better.
Index Terms: children's speech, speaker recognition, factor analysis, i-vector

1. Introduction

The employment of speaker recognition technology for children could be beneficial in several application areas, including child security and protection and for educational related purposes. Speaker recognition itself is divided into speaker verification, in which the user provides voice sample and a claim for her/his identity, and speaker identification, in which the user only provides voice samples. The objective for speaker verification engines is to verify the claim of the user, while for speaker identification the task is to find the closest user model to that particular voice sample. Each of these subcategories could divided further to the text-dependent and the text-independent modes of operation. In this research we focused on both speaker identification and verification for text-independent mode of operation.

Some of the most compelling applications of spoken language technology in education involve children, but computer recognition of children's speech is particularly difficult. This is due to the fact that children have large differences in both the acoustic and the linguistic aspects of speech compared to adults [1, 2, 3]. Differences in the pitch, the formant frequencies, the average phone duration, the speaking rate, the glottal flow parameters, pronunciation and grammar are the various acoustic and linguistic differences [4]. As reported, children have different values of mean and variance of the acoustic features of speech than those of adults [1]. For example, the area of the F1-F2 formant ellipses is larger for children than for adults for most vowel phonemes and children speech contains more dis-fluency and extraneous speech [5].

As all children undergo rapid development and with varying rates, it is difficult to model their constantly changing speech characteristics. Also as children grow, their speech production organs change and so their anatomy and physiology keep changing quite significantly. Thus, comparing with those of adults speech, children's speech has higher inter- and intra-speaker acoustic variability [2].

Compared to adults, children have formants located at higher parts of the spectrum [1, 4]. Also, they have high pitch frequency values which cause large spacing between the pitch harmonics [1, 4]. These high formant frequencies and pitch frequency values are attributed to their inherent shorter vocal tract and vocal folds lengths, respectively.

On the other hand, due to physiological differences between the speakers, the differences in the voice source parameters affect the source spectrum [6]. Contributing to all this is their increased intra-speaker spectral and temporal variability [7, 8, 9]. Increases in the intra-speaker spectral and temporal variability cause greater overlapping of the phonemic classes making the pattern classification problem even more difficult.

During the last decades different approaches have been examined for automatic speaker characterization. First attempts to tackle this problem date back to the 1970s [10]. To structure our discussion we can divide most of the deployed approaches into two main categories; phonotactic and acoustic approaches [11]. Acoustic approaches are the main focus of this work. These approaches does not need any specialized language knowledge. Acoustic based approaches can be applied to identify paralinguistic speaker characteristics. They have been widely used in different speaker characterization problems [11, 12, 13, 14]. In [15], Gaussian mixture model (GMM) mean super-vectors and SVM were applied. In the field of speaker recognition, recent advances using i-vectors have increased the recognition accuracy considerably [16]. A summary of the main approaches can be found in [11].

This paper structured as follows; Section 2 describes the speech corpus and the data is used for the experiments in this research. Section 3 contains the details of our speaker recognition systems. In Section 4 the experimental results are presented and analyzed, and finally Section 5 summarizes our findings.

2. The OGI kids speech corpus and data description

The OGI Kids Speech corpus ([17]) is used in all experiments. It contains examples of spontaneous and read speech from approximately 1100 children (roughly 100 children per grade from kindergarten (5–6 years old) to grade 10 (15–16 years old)), recorded in Portland, Oregon. Prompts were displayed as text on a screen, and a human recording of the prompt was played in synchrony with facial animation using the an-

10.21437/WOCCI.2017-10

imated 3D character Baldi. The subject repeated the prompt (read speech) or talked on a chosen topic (spontaneous speech). Recordings were made at 16 kHz sampling rate, 16 bits precision, using a head-mounted microphone. An average session lasted 20 minutes and yielded approximately 63 utterances (8-10 minutes of speech). However, for some subjects much less data is available. In speaker recognition experiments just 144 and 48 seconds of speech were used per subject for training and testing, respectively, for consistency between subjects. Table 1 and Figure 1 show the number of children recorded per grade and the distribution of children's ages, respectively. In the first column of Table 1 the blue, red, and black grades correspond to the age group (AG), labelled as AG1, AG2, and AG3. Choosing a AG banding was a big challenge. It was a trade-off between amount of data available for train and test, and level of variability in each AG. We decided to give the first priority to the amount of available data for training and testing, as it will lead to more statistically reliable result.

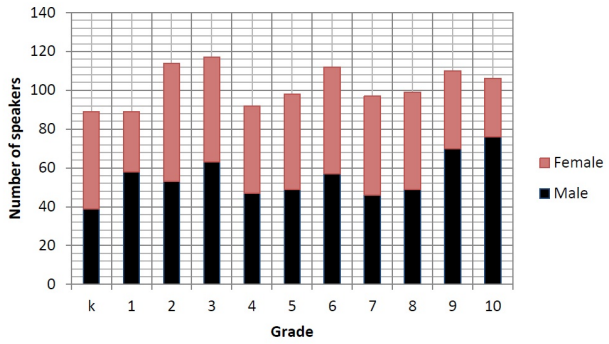

Figure 1: *Distribution of ages of children recorded in the CSLU children's speech corpus.*

Table 1: *Number of kids recorded for each grade.*

Grade	Age	# of speakers	
	(years)	Male	Female
K	5-6	39	50
1	6-7	58	31
2	7-8	53	61
3	8-9	63	54
4	9-10	47	45
5	10-11	49	49
6	11-12	57	55
7	12-13	46	51
8	13-14	49	50
9	14-15	70	40
10	15-16	76	30

Understanding how the performance of speaker recognition varies with age is another objective of this work. The grade that is likely to differ most from the others is K, because the speech of 5–6 year-olds may exhibit additional variability due to phonological factors associated with language acquisition ([18]). Hence the first age-group, AG1, should span as few grades as possible. For speaker recognition in our work AG1 comprises grades K, 1 and 2. Table 2 summarises the allocation of grades (age range) to age-group for the speaker recognition experiments.

Table 2: *Definitions of the age-groups used in our experiments. SR stands for Speaker Recognition.*

SR Experiment	AG1	AG2	AG3
Grade	K–2	3–6	7–10
Age	5–8 years old	9–12 years old	13–16 years old

Different experiments require different partitions of the corpus into training, development and test sets, to achieve a suitable balance between classes or to study a particular effect. In our experiments we have used 12 simulated classrooms of 30 children, 4 per each age group, and 2 simulated schools of 288 children, balanced across age and gender.

3. Speaker recognition systems

3.1. Signal analysis

Silence was discarded using energy-based speech activity detection. Frames of length 20 ms (10 ms overlap, Hamming window) were extended to 512 samples and a DFT was applied, giving a frequency resolution of 31.25 Hz. The resulting magnitude spectrum was passed to a bank of 24 Mel-spaced triangular filters, spanning frequencies from 0 Hz to 8000 Hz. For the speaker recognition experiments the outputs of all 24 filters were transformed into 19 static plus 19 Δ and 19 Δ^2 MFCCs.

3.2. Modeling

Speaker recognition systems which are used in this research are based on the factor analysis frame-work. In this approach i-vectors are used as the new sets of features. This assumes that a GMM supervector μ can be decomposed as $\mu = m + Tw$, where m is the UBM mean supervector, T is a linear mapping from a low-dimensional 'total variability subspace' W into the supervector space, and $w \in W$ is an 'i-vector', sampled from a standard normal distribution.

Two approaches to i-vector scoring were employed. In the first [19, 20], linear discriminant analysis (LDA) is applied to W, and each class c is represented by the mean m_c of length normalized i-vectors for that class in the LDA sub-space. At the recognition stage, the score for each class c is the dot product of the normalized LDA test i-vector with m_c (the cosine of angle between the test i-vector and m_c). This approach is referred to as 'i-vector'. A block diagram which shows this i-vector approach is depicted in Figure 2.

The second i-vector method, applied to speaker recognition, uses probabilistic LDA (PLDA) [21]. Before applying PLDA, the i-vectors are length-normalised and whitened [22]. Recognition is based on the log-likelihood ratio between same versus different speaker hypotheses [22]. This is referred to as 'i-vector-PLDA'.

Figure 3 shoes a block diagram which corresponds to the i-vector-PLDA approach.

In school experiments, the size of the T matrix and the i-vector dimension after LDA were set, empirically, to 400 and 300, respectively. In classroom experiments, the size of the T matrix and the i-vector dimension after LDA were set, empirically, to 400 and 40, respectively. The T matrix was learned using all training data.

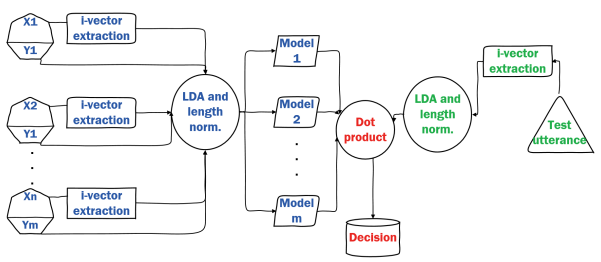

Figure 2: *A block diagram of the speaker recognition system based on the i-vector approach (scoring with cosine similarity measure), depicting both training and testing phase. X_n and Y_n represent samples and class labels, respectively.*

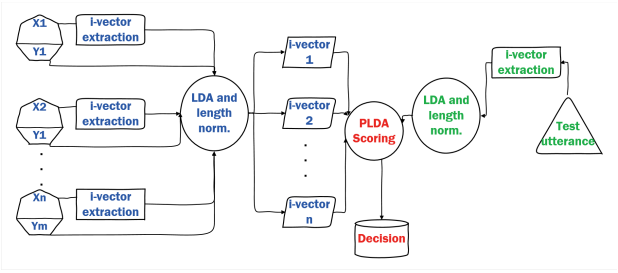

Figure 3: *A block diagram of the speaker recognition system based on the i-vector-PLDA approach, depicting both training and testing phase. X_n and Y_n represent samples and class labels, respectively.*

4. Experimental results and discussion

Speaker verification in our work followed the NIST methodology [23], whereby a test utterance is compared with eleven models, the "true" model plus ten randomly selected impostor models. However speaker identification is the classification problem of one out of total number of classes (which in this research classes are speaker, i.e. 30 for classroom and 288 for school simulated environments).

Table 3 shows identification and verification accuracy for a child in a simulated classroom or school, using i-vector-PLDA and 10 seconds test utterances. The experiment uses 12 "classrooms" of 30 children, 4 for each age-group, balanced across gender for classroom experiments and 2 simulated schools of 288 children, balanced across age and gender.

The results follow the expected pattern, with identification accuracy increasing and verification EER decreasing with age.

Table 4 shows identification and verification accuracy for a child in a simulated classroom or school, using i-vector (cosine similarity measure) and 10 seconds test utterances. The experiment uses 12 "classrooms" of 30 children, 4 for each age-group, balanced across gender for classroom experiments and 2 simulated schools of 288 children, balanced across age and gender. Same as i-vector PLDA approach the results follow the expected pattern, with identification accuracy increasing and verification EER decreasing with age.

Comparison of Table 3 and Table 4 reveals an interesting behaviour for these two scoring methods within i-vector framework. Obtained results show that PLDA scoring performs better for school experiments while the cosine similarity measure

Table 3: *Speaker verification EER and identification accuracy for three different age-groups (AG1, AG2 and AG3) obtained using the i-vector-PLDA system for the case of a child in a classroom and in a school scenario, respectively.*

Classroom scenario	Verification EER (%)	Identification Acc (%)
AG1 (K-2^{nd})	1.89	86.39
AG2 (3^{rd}-6^{th})	1.39	93.61
AG3 (7^{th}-10^{th})	0.97	97.34
School	1.04	86.89

scoring provide better performance for the classroom experiments. This suggest the usage of i-vector PLDA approach for problems with large number of classes involved (in here class refer to speaker) and the i-vector (using cosine similarity measure for scoring) for the problems with smaller number of classes,m e.g. classroom experiment in this research.

Table 4: *Speaker verification EER and identification accuracy for three different age-groups (AG1, AG2 and AG3) obtained using the i-vector system (with cosine similarity measure) for the case of a child in a classroom and school scenario.*

Classroom scenario	Verification EER (%)	Identification Acc (%)
AG1 (K-2^{nd})	1.94	88.79
AG2 (3^{rd}-6^{th})	1.18	94.76
AG3 (7^{th}-10^{th})	0.94	98.99
School	1.84	80.70

5. Conclusions

The objective of this paper is to evaluate the usage of two scoring methods within i-vector framework for possible practical applications of speaker recognition from children's speech. Scoring methods are applied to identify/verify speakers in classroom and school simulated environment. The classroom experiment uses 12 "classrooms" of 30 children, 4 for each age-group, balanced across gender, while for the school experiments 2 simulated schools of 288 children, balanced across age and gender were used. The scoring methods are based on PLDA and cosine similarity measure approaches.

Using the PLDA scoring approach for the speaker recognition experiment in the school environment results in 43.48% (relative) and 6.19% (absolute) improvements in the performance of speaker verification and identification tasks in term of reduction in EER and identification rate, compared with the cosine similarity approach, respectively. Classroom experiments shows opposite behaviour for almost all verification and identification experiments, and for all age-groups. This finding suggests the usage of i-vector PLDA approach for problems with large number of classes involved (in here class refer to speaker) whilst the i-vector (using cosine similarity measure for scoring) for the problems with smaller number of classes, e.g. the classroom experiment in this research.

6. Acknowledgements

This work was partially done as a part of first author's PhD research [24] and completed further recently. The authors would like to thank supportive members of school of Electronic, Electrical and Systems Engineering at University of Birmingham and specifically Professor Martin Russell for his supportive supervision.

7. References

[1] S. Lee, A. Potamianos, and S. 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.

[2] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 603–616, Nov 2003.

[3] A. H. M. R. S. Safavi, M. Najafian and P. Jancovic, "Comparison of speaker verification performance for adult and child speech," in *WOCCI*, 2014.

[4] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: duration, pitch and formants." in *EUROSPEECH*. ISCA, 1997. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/eurospeech1997.htmlLeePN97

[5] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, Oct 2007, pp. 22–25.

[6] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[7] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 2, pp. 65–78, Feb 2002.

[8] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 11, pp. 847 – 860, 2007, intrinsic Speech Variations.

[9] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analyzing children's speech an acoustic study of consonants and consonant-vowel transition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–393–I–396.

[10] R. Leonard, G. Doddington, and T. I. I. DALLAS., *Automatic Language Identification*. Defense Technical Information Center, 1974. [Online]. Available: http://books.google.co.uk/books?id=nlU8OAAACAAJ

[11] A. Hanani, M. Russell, and M. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Computer Speech Language*, vol. 27, no. 1, pp. 59–74, 2013.

[12] S. Safavi, A. Hanani, M. Russell, P. Jančovič, and M. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification." *IEEE Signal Processing Letters.*, vol. 19, no. 12, pp. 829–832, 2012. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=06317140

[13] S. Safavi, P. Jančovič, M. J. Russell, and M. J. Carey, "Identification of gender from children's speech by computers and humans." in *INTERSPEECH*. ISCA, 2013, pp. 2440–2444.

[14] M. Najafian, S. Safavi, P. Weber, and M. Russell, "Identification of british english regional accents using fusion of i-vector and multi-accent phonotactic systems," in *ODYSSEY*, 2016, pp. 132–139.

[15] S. Safavi, M. Najafian, A. Hanani, M. Russell, P. Jančovič, and M. Carey, "Speaker recognition for children's speech," *Interspeech*, pp. 1836–1839, 2012.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[17] K. Shobaki, J. P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers," *Int. Conf. on Spoken Language Processing*, 2000.

[18] E. Fringi, J. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," 2015, pp. 1621–1624.

[19] E. Singer, P. A. Torres-Carrasquillo, A. Reynolds, D. A. McCree, N. Richardson, F. amd Dehak, and D. E. Sturim, "The MITLL NIST LRE 2011 language recognition system." ISCA, 2012, pp. 209–2015.

[20] S. Safavi, M. J. Russell, and P. Jančovič, "Identification of age-group from children's speech by computers and humans." in *INTERSPEECH*. ISCA, 2013, pp. 2440–2444.

[21] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE Int. Conf. on Computer Vision*, 2007, pp. 1–8.

[22] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," 2011, pp. 249–252.

[23] N. M. I. Group., "2003 nist speaker recognition evaluation ldc2010s03." in *Philadelphia: Linguistic Data Consortium*, 2010.

[24] S. Safavi, "Speaker characterization using adult and childrens speech," Ph.D. dissertation, University of Birmingham, 2015.