Assessing the effectiveness of diarization algorithms in costa rican children-adult speech according to age group and gender

Evaluación de la efectividad de los algoritmos de registro en el habla de niños y adultos costarricenses según grupo de edad y género

Alejandro Chacón-Vargas[1], Daniel Pérez-Conejo[2], Marvin Coto-Jiménez[3]

Chacón-Vargas, A; Pérez-Conejo, D; Coto-Jiménez, M. Assessing the effectiveness of diarization algorithms in costa rican children-adult speech according to age group and gender. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 24-32.

https://doi.org/10.18845/tm.v35i8.6443

[1] University of Costa Rica. Costa Rica. E-mail: alejandro.chaconvargas@ucr.ac.cr

[2] University of Costa Rica. Costa Rica. E-mail: daniel.perezconejo@ucr.ac.cr

[3] University of Costa Rica. Costa Rica. E-mail: marvin.coto@ucr.ac.cr
https://orcid.org/0000-0002-6833-9938

Keywords

Children's speech; clustering; speaker diarization; speech processing.

Abstract

Speaker diarization is the task of automatically identifying speaker identities and detecting their speaking times in an audio recording. Several algorithms have shown improvements in the performance of this task during the past years. However, it still has performance challenges in interaction scenarios, such as between a child and adult, where interruptions, fillers, laughs and other elements may affect the detection and clustering of the segments.

In this work, we perform an exploratory study with two diarization algorithms in children-adult interactions within a recording studio and assess the effectiveness of the algorithms in different age groups and genders. All participants are native Costa Rican Spanish speakers. The children have ages between 3 to 14 years, and the interaction combines guided repetition of words or short phrases, as well as natural speech.

The results demonstrate how the age affects the diarization performance, both in cluster purity and speaker purity, in a direct but non-linear fashion.

Palabras clave

Habla de los niños; agrupación; registro del hablante; procesamiento del habla.

Resumen

El registro de los oradores es la tarea de identificar automáticamente las identidades de los oradores y detectar sus tiempos de conversación en una grabación de audio. Varios algoritmos han mostrado mejoras en el desempeño de esta tarea durante los últimos años. Sin embargo, todavía presenta desafíos de desempeño en escenarios de interacción, como entre un niño y un adulto, donde las interrupciones, los rellenos, las risas y otros elementos pueden afectar la detección y agrupamiento de los segmentos.

En este trabajo, realizamos un estudio exploratorio con dos algoritmos de registro en interacciones niños-adultos dentro de un estudio de grabación y evaluamos la efectividad de los algoritmos en diferentes grupos de edad y géneros. Todos los participantes son hispanohablantes nativos de Costa Rica. Los niños tienen edades comprendidas entre los 3 y los 14 años, y la interacción combina la repetición guiada de palabras o frases cortas, así como el habla natural.

Los resultados demuestran cómo la edad afecta el rendimiento del registro, tanto en la pureza del grupo como en la pureza del hablante, de forma directa pero no lineal.

Introduction

Speaker diarization is often described as the task of automatically responding to the question of “who spokes when?” in an audio recording. The answer to the question, performed using a diversity of algorithms and approaches, is the time periods when each speaker is active, even for cases where the number of speakers is unknown.

Determining those time periods for each speaker can be useful in tasks, such as speaker indexing and the retrieval of large audio sets. Even transcription in movies, where automatically adding punctuation and speaker markers is of interest due to the large volume of films produced and the possibility of making them accessible for people with disabilities [1]. Also, the segmentation produced with the time marks can benefit automatic speech recognition of more homogeneous segments, with a single speaker in each one.

The most typical scenarios are meetings, broadcast news and conversational telephone speech. In those cases, various problems need to be solved, including the building of clusters of information to estimate the number of speakers, and most of the times without a priori information, the processing of audios with music, silence, and other sounds, and the processing of spontaneous speech with overlapping voices of speakers [2].

More recently, the number of applications involving new scenarios and conditions have renewed interest in this task [3]; for example, the interaction between children and adults in clinical assessments and evaluations for children with autism spectrum disorder.

One of the reasons for the increasing interest of the research community in this topic is the availability of multimedia information in recent years. With this increase in the amount of information, more demanding conditions arose, for example multi-genre data, multiple speakers, diverse acoustic conditions or multiple recordings. A way to assess the progress and usefulness of this techniques is through evaluations in a particular domain [1].

Compared to the number of studies made for improving the performance of diarization systems, few studies have systematically analyzed the different sources and types of diarization errors [3].

In this work, we build an annotated corpus of children-adult interaction within a recording studio, and explore two diarization algorithms to determine systematically the performance and source of errors, especially those derived from the age range and gender of the children. To the best of our knowledge, this is the first study of speaker diarization for a population of Costa Rican children, which includes both guided repetition of words and phrases, and natural interactions.

As previous studies have reported in other languages, it is difficult for a child to focus his or her attention for long periods of time or remain in the same place. Thus, it is normal that the recording includes noises and interactions between the child and objects around [4]. For this reason, the diarization of children's speech is a particularly challenging task.

Related work

Among the many references on speaker diarization algorithms and conditions, in this section we describe those more closely related to the challenging conditions, and in particular the speech/adult interactions. For example, a recent experiment at the Johns Hopkins University [5] explored a variety of difficult conditions that demonstrate a high variability in the results, according to those conditions and the selection of features and procedures. Speaker overlap is an extra challenging condition for the algorithms [6], and is frequently found in recordings with children.

Most diarization methods take short segments of an audio recording and group the segments based on the clustering of a vector of parameters extracted from each one to conform to an i-vector (or fixed size vector). For example, such a procedure has been followed in [4] for the case of children's speech, using a Probabilistic Linear Discriminative Analysis with publicly available data for the English language. Some recent proposals have also considered additional elements of the recordings in order to improve the results, in English language [3]. Diarization of children and adult interaction for the English Language have been recently analyzed in [7]. In this case, the main concern is the analysis of the speech of children with an autism diagnosis. This kind of interaction is referred to as child-adult speaker classification in conversations with specified roles. In our work, a similar task is done for Costa Rican children, with the additional analysis of age and gender dependency.

Additionally, from collecting speech for speech technologies (e.g., speech syn- thesis and recognition) the procedure of diarization is of interest to language development researchers [8]. For example, determining the content of the child's language environment using a wearable recording unit, and categorizing the results according to whom initiated the speech and the number of interactions [9].

The challenges in the diarization of children's speech have also seen reported in [10]. According to the authors, children's recordings may contain speech un- familiar to usual processing models, such as cooing or crying. These kinds of elements can affect the algorithms that have proven efficiency with pure speech data. Additionally, the reference report that adults' speech when talking to infants has a larger pitch range.

Another factor that affects the accuracy of algorithms is the acoustic quality of the audio recording, distance between speaker and microphone and back- ground noise. The accuracy of previous references using a single microphone located 1-2m from the speakers, achieved an accuracy rate of 70% [11].

In this work, we perform a first study on the diarization of children-adult guided interactions, using a state-of-the art algorithm, with the purpose of verifying the conditions in the literature that affects the algorithms in this task, for the particular case of Costa Rican children. As we are focused on obtaining quality information from children of all ages, we classify the results according to that parameter and also the gender, to determine whether or not our interaction strategies are producing recordings suitable for automatic analysis.

The rest of this paper is organized as follows: Section 2 presents the experimental setup. Section 3 describes the results and discussion, and finally Section 4 presents the conclusions and future work.

Experimental Setup

In this section we summarize the main elements for building the database and the experiments performed to assess the diarization algorithms.

Database Recording and Processing

To obtain the recordings of the children-adult interactions, we conducted several recording sessions at a professional studio. The recordings were carried out using professional equipment, with one child and one adult each time. Those recordings were then manually edited and carefully annotated in terms of speaker turns and times, to define the ground truth for testing the algorithms.

Children Interaction

We established a pre-defined strategy with the use of pictograms designed to obtain repetitions of words with the children's voice. These recordings had a length of about 15 minutes for the children aged 8 to 14 years old, and about 10 minutes for children in the range of 3 to 7 years. Due to the limited time of concentration of some children, the interaction also considered some segments of jokes and laughs. Such elements are present more frequently in the recording of the younger children.

Diarization Algorithms

The Fisher Discriminant Analysis (FLD) is an algorithm for grouping vector of features. The criteria of the method is that the classes' means of each group are separated, and the variance within each group is small. According to the method describe in [12], a set of matrices is build, considering the mean of all vector of features *m* and the mean of all vector of features of each class m_c.

For example, the class scatter matrix is defined as

$$S_b = \sum_{c\in C}[(m_c - m)(m_c - m)^T] \quad (1)$$

where *C* is the set of all classes.

The average within-class scatter matrix S_v is defined as

$$S_v = \sum_{c\in C}\sum_{x\in C}[(x - m_c)(x - m_c)^T] \quad (2)$$

where *x* are the feature vectors.

The total scatter matrix of samples is defined as

$$S_m = \sum_{allx}[(x - m)(x - m)^T] \quad (3)$$

The aim of FLD is to find a matrix A that optimize a criterion of grouping separation in an optimal subspace.

The case of Fisher Semi Discriminant Analysis is applied when the set of classes are unknown. The idea, presented by [13] is to assume that neighbor samples of an audio file are likely to be part of the same class.

Evaluation

According to previous references [13, 14], the common measures for the effective- ness of a speaker diarization are the cluster purity and speaker purity, defined as follows:

- ***Cluster purity:*** The cluster purity is the percentage of data in each cluster which belongs to the most dominant speaker, according to the ground-truth (annotated) reference. Mathematically, it can be expressed as:

$$\text{Cluster purity} = \frac{1}{N}\sum_{i=1}^{N_C}\max_{j=1,\ldots,N_S} n_{ij}, \quad (4)$$

where *N* is the total number of detected segments, N_s is the number of speakers, N_c is the total number of clusters, and n_{ij} is the total number of segments classified in cluster *i* and spoken by speaker *j*.

- ***Speaker purity:*** The speaker purity is the percentage of data of the most common detected speaker within each speaker class. This is a measure for the assessment of the speaker turns, and can be expressed as:

$$\text{Speaker purity} = \frac{1}{N}\sum_{i=1}^{N_S}\max_{j=1,\ldots,N_C} n_{ij}, \quad (5)$$

where each symbol follows the previous description.

Results and Discussion

This section introduces the results of the diarization algorithms for the children speaking sessions in the developed database. The results are organized comparing FLSD algorithm with the simplest method of K-means.

In regards to cluster purity, the results in relation to participants' age are shown in table [1] for K-means algorithm. It is noted that it tends to the increase of the cluster purity as children age progresses. In some cases, they will increase or decrease, but they always tend to increase. Equally, Speaker Purity tendency tends to increase with ages, where it increases or decreases.

A similar trend is noted in Table [2] for the FLSD algorithm. In the case of Cluster Purity, a total of seven age ranges presents better results than the K-means case. As to Speaker Purity, the comparison with k-means case is for improvement: there are a total of eight cases that give a better result. Among the reasons that can be pointed out for the decrease of K-means and FLSD algorithms' capacity for carrying out diarization is that there are more elements of noise, short interactions, interruptions, and other speak fillers in younger children's interaction.

A comparison of the results according to children's gender is observed in graphics [1] and [2]. A tendency to obtain better results in boys than in girls is observed. This can be partly explained by the fact that the adult person who interacts with the children in all sessions is a woman; thus, it is more likely that there is a higher contrast with the boys' voices that allow a high contrast for the algorithm.

Table 1. Diarization results for K-means.

Session	Age	Number of speaking turns	Cluster purity	Speaker purity
1	3	120	84.4	51.7
2	4	125	84.4	51.7
3	6	117	83.3	73.7
4	6	103	74.0	74.0
5	7	101	72.7	72.7
6	7	114	72.9	57.6
7	8	96	79.3	79.6
8	11	144	72.4	59.1
9	11	68	80.2	73.2
10	12	170	72.3	50.0
11	12	53	83.6	83.6
12	14	49	98.2	97.4

Table 2. Diarization results for FLSD.

Session	Age	Number of speaking turns	Cluster purity	Speaker purity
1	3	120	84.4	41.6
2	4	125	84.4	41.6
3	6	117	74.3	74.3
4	6	103	83.3	82.5
5	7	101	79.2	79.2
6	7	114	72.9	70.8
7	8	96	79.3	79.0
8	11	144	77.8	77.8
9	11	68	80.2	73.7
10	12	170	72.8	52.1
11	12	53	85.1	85.1
12	14	49	98.5	95.0

As is observed in these results, there is a clear dependency of diarization algorithm capacity, according to the participants' age and gender. The use of these algorithms for this application environment does not seem to have dependable levels, so higher explorations must be realized, especially for the cases of younger female children.

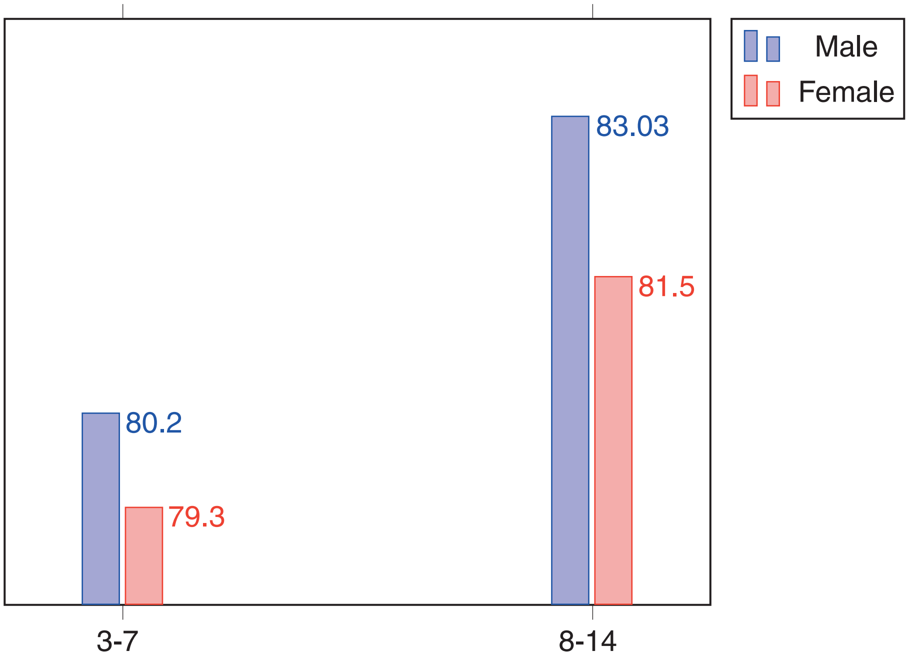

Figure 1. Cluster purity by gender (FLSD algorithm)

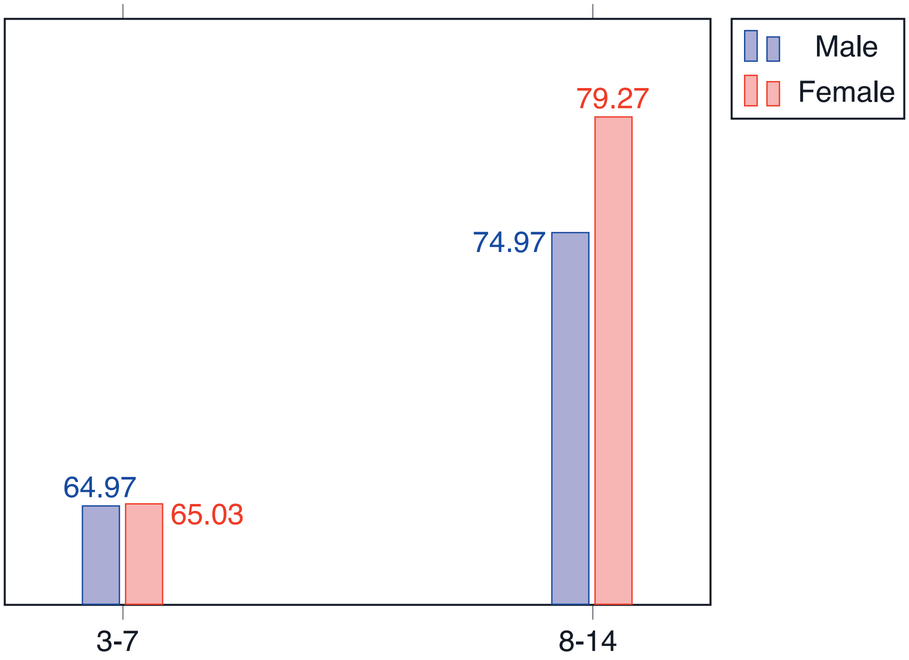

Figure 2. Speaker purity by gender (FLSD algorithm)

Conclusions

The purpose of this research is to determine the effectiveness of two diarization algorithms in relation to the age and gender of children. We perform the analysis in children-adult interactions with native Costa Rican Spanish speakers.

The results show that there is a direct correlation between the effectiveness of the algorithm and the age and gender of children. According to age, both K-means and FLSD algorithms tended to perform better in older children, and in relation to gender, male children obtained better results.

The results presented here, open up the possibility of exploring new specific strategies of diarization according to the age and gender of children, in order to improve the segmentation and grouping processes of speech in this type of sessions.

In regards to future work, it is intended to debug existing algorithms to improve their performance with young children, new accents, and noisy environments. It can also be explored with a greater combination of algorithms to obtain higher levels of reliability.

References

[1] Karanasou, Penny, et al. ”Speaker diarization and longitudinal linking in multi- genre broadcast data.” 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.

[2] Meignier, Sylvain, et al. ”Step-by-step and integrated approaches in broadcast news speaker diarization.” Computer Speech & Language 20.2-3 (2006): 303-330.

[3] Kumar, Manoj, et al. ”Improving speaker diarization for naturalistic child-adult conversational interactions using contextual information.” The Journal of the Acoustical Society of America 147.2 (2020): EL196-EL200.

[4] Xie, Jiamin, et al. ”Multi-PLDA Diarization on Children's Speech.” Interspeech. 2019.

[5] Sell, Gregory, et al. ”Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge.” Interspeech. 2018.

[6] Fujita, YusukeRao, et al. ”Meta-Learning for Robust Child-Adult Classification from Speech.” ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[7] Koluguri, Nithin Rao, et al. ”Meta-Learning for Robust Child-Adult Classification from Speech.” ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[8] Najafian, Maryam, and John HL Hansen. ”Speaker independent diarization for child language environment analysis using deep neural networks.” 2016 IEEE Spo- ken Language Technology Workshop (SLT). IEEE, 2016.

[9] Zhou, Tianyan, et al. ”Speaker diarization system for autism children’s real-life audio data.” 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2016.

[10] Karadayi, Julien, Camila Scaff, and Alejandrina Cristià. ”Diarization in Maximally Ecological Recordings: Data from Tsimane Children.” SLTU. 2018.

[11] Gorodetski, Alex, Ilan Dinstein, and Yaniv Zigel. ”Speaker diarization during noisy clinical diagnoses of autism.” 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019.

[12] Sarafianos, Nikolaos, Theodoros Giannakopoulos, and Sergios Petridis. ”Audio- visual speaker diarization using fisher linear semi-discriminant analysis.” Multimedia Tools and Applications 75.1 (2016): 115-130.

[13] Giannakopoulos, Theodoros, and Sergios Petridis. ”Fisher linear semi-discriminant analysis for speaker diarization.” IEEE transactions on audio, speech, and language processing 20.7 (2012): 1913-1922.

[14] Chen, Liping, et al. ”On Early-stop Clustering for Speaker Diarization.” Proc. Odyssey 2020 The Speaker and Language Recognition Workshop. 2020.