A first study on age classification of costa rican speakers based on acoustic vowel analysis

Un primer estudio sobre clasificación por edades de hablantes de costarricense basado en análisis de vocales acústicas

Victor Yeom-Song[1], Marvin Coto-Jiménez[2]

Yeom-Song, V.; Coto-Jiménez, M. A first study on age classification of costa rican speakers based on acoustic vowel analysis. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 145-152.

https://doi.org/10.18845/tm.v35i8.6466

[1] University of Costa Rica. Costa Rica. E-mail: victor.yeom@ucr.ac.cr
https://orcid.org/0000-0003-4172-1536

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

Keywords

Age recognition; children's speech; classification; vowel analysis.

Abstract

According to several studies, children's speech is more dynamic and inconsistent compared to an adult's speech. This aspect can be considered in the task of recognizing the age of the person who speaks and of great importance in many applications, such as human-computer interaction, security on Internet and education assistants. Those applications have a dependency on language and accent, due to the different sounds and styles that characterize the speakers. This paper presents the initial results on the identification of Costa Rican children's speech, in a database created for this purpose, consisting of words pronounced by adults and children of several ages. For this first study we chose the most common vowel of the language, and extract a set of common acoustic features to determine its applicability in distinguishing between adults and children of an age range. The outcome results shows promising results in the classification using a single vowel, that improves according to the number of vowels used to extract the acoustic features. This means that an automatic system could be able to improve its capacity to identify age as more speech information is received and transcribed, but cannot be very accurate in short interactions.

Palabras clave

Reconocimiento de edad; habla infantil; clasificación; análisis de vocales.

Resumen

Según varios estudios, el habla de los niños es más dinámica e inconsistente en comparación con el habla de un adulto. Este aspecto se puede considerar en la tarea de reconocer la edad de la persona que habla y de gran importancia en muchas aplicaciones, como la interacción humano-computadora, la seguridad en Internet y los asistentes educativos. Esas aplicaciones tienen una dependencia del lenguaje y el acento, debido a los diferentes sonidos y estilos que caracterizan a los hablantes. Este trabajo presenta los resultados iniciales sobre la identificación del habla infantil costarricense, en una base de datos creada para tal fin, que consta de palabras pronunciadas por adultos y niños de distintas edades. Para este primer estudio, elegimos la vocal más común del idioma y extraemos un conjunto de características acústicas comunes para determinar su aplicabilidad para distinguir entre adultos y niños de un rango de edad. Los resultados obtenidos muestran resultados prometedores en la clasificación utilizando una sola vocal, que mejora según el número de vocales utilizadas para extraer las características acústicas. Esto significa que un sistema automático podría mejorar su capacidad para identificar la edad a medida que se recibe y transcribe más información del habla, pero no puede ser muy preciso en interacciones breves.

Introduction

Speech signals contain information of several kinds. The most visible part is the linguistic content but, additionally, there exists paralinguistic information associated with the message, such as the speaker's accent, gender, age, or emotional state [1]. Even the state of health, including those ones related to the recent COVID-19 outbreak, have been successfully investigated using speech signals [2] or sound associated with the vocal tract [3].

Research in the processing of such paralinguistic information has grown considerably over the last two decades, including, more recently, children's speech [4]. One particular task of this processing is the age and gender recognition from speech recordings.

Automatic recognition of the paralinguistic information that allows the identification of children from their speech can be of benefit in many application areas. For example, to guide a child-computer (or child-robot) interaction, automatically adapt content, enhance child security in communications and inter- actions through the Internet, and educational applications [4].

Most current Automatic Speech Recognition (ASR) systems are not particularly accurate with children's speech, especially those in preschool age [5]. The main explanation for this issue is the acoustic mismatch between children's speech and the information used to train the recognizers. To overcome this problem it is vital to analyze and describe the characteristics of children's speech in each language, and develop systems that can adapt in terms of recognizing and, in the case of two way communication, even adapt its vocabulary when a child is speaking.

As children grow, these automatic recognition rate improve, as well as the general understanding of the words for native speakers [6], which can be related to the closer characteristics of the sounds and general articulation during the years of development.

Those characteristics of children's speech vary rapidly as a function of age, due to the anatomical and physiological changes occurring during their development [7]. Although a quantitative study of a cross lingual validation of this statement has not been addressed, it is clear that the common changes during a child's development affects and improves the acoustical characteristics of their speech.

Acoustic analysis of children speech, especially for younger ages, is a challenging task from the very beginning of data recording, where an interaction strategy should be applied to establish a proper environment for the participation of children. But, in terms of developing human-computer interaction for this population, including speech recognition, it is mandatory to establish ways of automatic identification of children using their voices' acoustic characteristics. For this reason, in this work we conducted a first study on Costa Rican speakers' age classification, particularly in two classes: children and adults. Our proposal seeks to explore the possibility of performing this classification using acoustic characteristics of the most common vowel in the Spanish language.

Related work

The analysis of formants, pitch and duration in the English language can be tracked back [8] for children of 5 to 18 years old. This study also pioneered the analysis of vowel characteristics for this population in the English language, combined with duration, and spectral variations.

The study of vowels in several languages in terms of recognition of sounds has been presented in [9] with a recognition rate of 76.25%. In the Russian language, the vowels pronounced by children of 6 and 7 years have been analyzed in [6]. The main differences between the vowels of this population and Russian adults are longer duration and high pitches. Also, the formant structure is not formed completely in children of these age ranges.

For the case of the English language, a characterization of children's vowels in terms of variability in repetitions has been presented in [7], with higher variability in the younger population. As this study shows, the majority of the analysis of children's speech dealt with vowel duration, pitch, and formants. The analysis of consonants is less common in the literature. Other acoustic characteristics, such as co-articulation, have also been studied in [10].

Also for the English language, in [5], acoustic characteristics of young children's speech were studied as a function of age. The fundamental frequency, formants and vowel duration for vowels were found to show age-dependent trends; particularly the variability of those parameters, with less significant changes in pitch.

For the Italian language, changes in acoustic characteristics of children have also been presented in [11], confirming that characteristics of children's speech change with age, and that spectral and temporal variability decrease as age increases. The variability shows substantial differences in children at an early age compared to children at a late age.

In terms of classification of children and adult speech, recent works have made use of pitch and formants with spectral coefficients [12]. The accuracy of the classification was above 97%. In the Italian language, recent experiences have reported accuracy higher than 80% [13] in children and adult recognition.

For the Portuguese language, a classifier comparison for children and adult speech recognition using Perceptual Linear Prediction coefficients and pitch as features were presented in [14]. Using a small amount of training data the accuracy of the best classifiers were as high as 97.4%.

In our work, we extend the application of acoustic analysis of vowels commonly performed in other languages for Costa Rican children, and compare some classifiers' performance to detect the voice of an adult or a child using only the acoustic analysis of such sounds.

The rest of this paper is organized as follows: Section 2 presents the experimental setup of our work. Section 3 shows the Results and Discussion, and finally in Section 4 we present the conclusions and future work.

Experimental Setup

Database

Two sessions for recording isolated words of children with ages between 4 to 12 years old were conducted, with a strategy of interaction with pictograms. These pictograms included simple words according to the development of the language of children in the first years. The same set of words was repeated for all the children, and a group of 8 adults. All the participants are native Costa Rican Spanish speakers.

The children were divided into two groups: Children in early childhood (those between within ages 4 to 7 years old), and Children in late childhood (those between ages 8 to 12 years old).

The selection of the same words of all the participants was manually edited and then segmented using the Praat system [15]. In this system, the temporal marks for each sound and its corresponding features can be extracted. For our experiments, we began with the selection of information of the five vowels of Spanish language: /a/, /e/, / i /, /o/, /u/.

To build the database for the experimentation with classifiers, we selected the most common vowel of the Spanish language: /e/ [16]. Then, we extracted subsets of three, five and ten randomly selected vowels within each age group of one single speaker each time. With this procedure, we pretend to emulate the vocal emission of any word or phrase, where it is not possible to establish prior information on the upcoming vowels and its characteristics.

Features

For each vowel, we extract the most common acoustic analysis, according to previous references, from sets of 3, 5 and 10 vowels: Average fundamental frequency (f0) in Hz, Minimum f0 in Hz, Maximum f0 in Hz, Average duration in seconds, Minimum duration in seconds, Maximum duration in seconds, Average position of the first formant of the spectrum (F1) in Hz, Minimum and Maximum of F1 in Hz, and the same measures of the second and third formants (F2 and F3). Detailed description of the algorithms applied in the features extraction can be found in [17].

Results and Discussion

In this section, we present the results obtained for the classification task with three different classifiers: Random Forest, Support Vector Machines (SVM), and k-Nearest Neighbors (kNN). It is to be noted that for kNN, we used k=1, so it is just a Nearest Neighbor classifier. We can see the classification results using 3 vowels in Table 1. We can see that, in this case, the Random Forest has the best classification performance overall, even though the scores, in general, are really similar for each classification task according to the age ranges. The best case out of all of the ones presented belongs to the Random Forest classifier classifying children in early childhood and adults, which is to be expected since the acoustic features under study present the most variability within these age groups.

The classification results from using 5 vowels are shown in Table 2. There is an evident increase in all the metrics across the board, but it should be noted that the Random Forest classifier for early children and adults had the least amount of improvement. The best classification case in general still belongs to the Random Forest for classification between children in late childhood and adults, even as the SVM has the best performance in the two other cases. This outcome could just be a result of the random seed used for the Random Forest, and may not be entirely indicative of Random Forest being completely better than the other classifiers.

It should be noted that with five vowels, each classifier’s performance is really similar for each respective case, with pretty much less than a 2% variation in accuracy for most cases. Thus, it really could not be said that one classifier is definitely better than another one in this task.

Table 1. Classification using data from three /e/ vowels.

Random Forest				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	90.00	0.90	0.90	0.90
Late childhood	85.00	0.85	0.85	0.85
All	87.10	0.87	0.87	0.87
SVM				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	86.67	0.87	0.87	0.87
Late childhood	83.89	0.85	0.84	0.84
All	87.10	0.87	0.87	0.87
kNN				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	85.00	0.85	0.85	0.85
Late childhood	83.33	0.84	0.83	0.83
All	86.13	0.87	0.86	0.86

Table 2. Classification using data from five /e/ vowels.

Random Forest				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	91.67	0.92	0.92	0.92
Late childhood	94.97	0.95	0.95	0.95
All	91.29	0.91	0.91	0.91
SVM				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	92.50	0.93	0.93	0.93
Late childhood	91.06	0.91	0.91	0.91
All	92.26	0.92	0.92	0.92
kNN				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	90.83	0.91	0.91	0.91
Late childhood	92.74	0.93	0.93	0.93
All	90.00	0.90	0.90	0.90

The classification results using 10 vowels are shown in Table 3. We can see that, in general, the results are even better than with 5 vowels, and that now the best classification performance belongs to the SVM in classification between early children and adults, with an astounding 99% accuracy. In other cases, other classifiers have better performance. Again, the difference in the results between one case and another differs by less than 2%, so there would not be one clear winner for classification.

Table 3. Classification using data from ten /e/ vowels.

Random Forest				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	98.33	0.98	0.98	0.98
Late childhood	97.22	0.97	0.97	0.97
All	97.74	0.98	0.98	0.98
SVM				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	99.17	0.99	0.99	0.99
Late childhood	96.11	0.96	0.96	0.96
All	95.81	0.96	0.96	0.96
kNN				
Age range	Accuracy (%)	Precision	Recall	F-measure
Early childhood	98.33	0.98	0.98	0.98
Late childhood	98.33	0.98	0.98	0.98
All	97.10	0.97	0.97	0.97

One aspect to be noted with the results is that in each case, all the metrics for accuracy, precision, and recall (and since the latter 2 are equal, F-measure as well) are numerically pretty much the same. This tells us that in each case the rate of false positives is equal to the rate of false negatives, which, in turn, means that the classification task is equally prone to type I and type II errors.

With the presented results, it can be seen that the classification of speech for children and adults is a very manageable task. In fact, with 5 and 10 vowels, the classifiers produced similar results between one another, with less than 2% variation in the used metrics. Since three different classifiers with reasonably different architectures managed to produce such similar results, it could be an indicator of possible ease in the classification stage.

As expected, the better classification results are obtained with 10 vowels as summarized in Figure 1, but the results with 3 and 5 vowels are really good as well. Since waiting for 10 utterances of a single vowel could be considered slow for real-life use, one could look at the previous results and consider using them depending on the application.

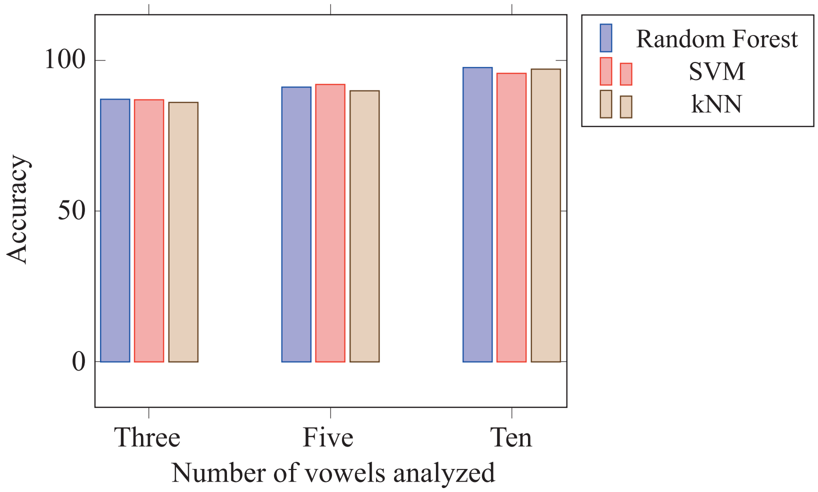

Figure 1. Results for each classifier for children in all age range and the amount of vowels analyzed.

Conclusions

The classification of speech between children and adults is a task with many potential applications, ranging from recommender systems to web security. As such, a study on the nature of the task itself is useful to explore its viability in implementation. This work offers a first approach to such a type of research for Costa Rican Spanish, using three different classifiers to assess the possibility of the task within a a framework of acoustic features.

Using only the vowel /e/ (the most frequent in the Spanish language), the results obtained are promising, considering that with only 3, 5, or 10 vowels produced a considerably high performance across the experiments, with over 80% accuracy using only 3 vowels. In general, the best results were obtained with 10 vowels. However, this doesn't necessarily mean that for real-life use, this is the best way to classify voice, as it could be considered slow depending on the application.

As for the task itself, the use of three different classifier types shows that it is quite manageable since all the classifiers produced really similar results in each study case. Thus, the implementation of a real-time speech classifier wouldn't be too far-fetched with current technology. Further exploration of the nature of the data could show better results in the future.

For future work, it is possible to consider the use of combinations of vowels for the task of classification, to study better classification boundaries or if.

Additionally, the application of other classifiers or set of parameters to achieve better results and contemplating the possibility of a real time age recognition system for Costa Rican Spanish.

References

[1] Safavi, Saeid, Martin Russell, and Peter Janˇcoviˇc. "Automatic speaker, age-group and gender identification from children's speech." Computer Speech & Language 50 (2018): 141-156.

[2] Schuller, Bjorn W., et al. "Covid-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 Corona crisis." arXiv preprint arXiv:2003.11117 (2020).

[3] Imran, Ali, et al. "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app." Informatics in Medicine Unlocked (2020)

[4] Safavi, Saeid, et al. "Identification of gender from children's speech by computers and humans." INTERSPEECH. 2013.

[5] Yildirim, Serdar, et al. "Acoustic analysis of preschool children's speech." Proc.15th ICPhS. 2003.

[6] Lyakso, Elena E., Olga V. Frolova, and Aleks S. Grigoriev. "The acoustic char- acteristics of Russian vowels in children of 6 and 7 years of age." Tenth Annual Conference of the International Speech Communication Association. 2009.

[7] Gerosa, Matteo, et al. "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition." 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Vol. 1. IEEE, 2006.

[8] Lee, Sungbok, Alexandros Potamianos, and Shrikanth Narayanan. "Analysis of children's speech: Duration, pitch and formants." Fifth European Conference on Speech Communication and Technology. 1997.

[9] Ting, Hua Nong, and Jasmy Yunus. "Speaker-independent Malay vowel recognition of children using multi-layer perceptron." 2004 IEEE Region 10 Conference TENCON 2004. IEEE, 2004.

[10] Katz, William F., and Sneha Bharadwaj. "Coarticulation in fricative-vowel syllables produced by children and adults: A preliminary report." Clinical linguistics & phonetics 15.1-2 (2001): 139-143.

[11] Gerosa, Matteo, Diego Giuliani, and Fabio Brugnara. "Acoustic variability and automatic recognition of children's speech." Speech Communication 49.10-11 (2007): 847-860.

[12] Zeng, Yumin, and Yi Zhang. "Robust children and adults speech classification."Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007). Vol. 4. IEEE, 2007.

[13] Massarente, Enrico. "Classificazione automatica della voce in ambito logopedico: training e testing di un algoritmo per discriminare la voce adulta da quella dei bambini." (2015).

[14] Martins, Rui, et al. "Detection of Children's Voices." I Iberian SLTech 2009: 77.

[15] Goldman, Jean-Philippe. "EasyAlign: an automatic phonetic alignment tool un- der Praat." Interspeech'11, 12th Annual Conference of the International Speech Communication Association. 2011.

[16] Guirao, Miguelina, and Mar´ıa Garc´ıa Jurado. "Frequency of occurence of phonemes in American Spanish." Revue quebecoise de linguistique 19.2 (1990): 135-149.

[17] Boersma, P. Weenink, D. "Praat: doing phonetics by computer" [Computer pro- gram]. Version 6.0.37, retrieved May 2020 from http://www.praat.org/.