

# Assessing costa rican children speech recognition by humans and machines


## Evaluación del reconocimiento de voz de los niños costarricenses por humanos y máquinas


Maribel Morales-Rodríguez<sup>1</sup>, Marvin Coto-Jiménez<sup>2</sup>

---

Morales-Rodríguez, M.; Coto-Jiménez, M. E. Assessing costa rican children speech recognition by humans and machines. *Tecnología en Marcha*. Vol. 35, special issue. November, 2022. International Work Conference on Bioinspired Intelligence. Pág. 74-82.

 <https://doi.org/10.18845/tm.v35i8.6453>

1 University of Costa Rica. Costa Rica.  
E-mail: [maribel.moralesrodriguez@ucr.ac.cr](mailto:maribel.moralesrodriguez@ucr.ac.cr)  
 <https://orcid.org/0000-0002-3426-5192>

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

## Keywords

Children speech; speech recognition; speech technologies; WER.

## Abstract

In recent years, an increasing number of studies on human-computer interaction is taking place, due to the pervasive speech interfaces implemented in systems such as cell phones, personal and home automation assistants. These studies include automatic speech recognition (ASR) and speech synthesis, and are considering a wider variety of conditions of the signals, such as noise and reverberation, and accents and age-related effects as well. For example, one of the key challenges is the development of ASR for children's speech. Since the current systems have a dependency on language and accents, thus, to improve it, the investigations of speech recognition technologies suitable for children are needed. In this paper, we assess commercial ASR systems for the recognition of Costa Rican children's speech, for users with ages ranging between three and fourteen years old. To establish a comparison and numeric validation of the ASR systems in recognizing children's isolated words, we conducted a large subjective listening test that computes the differences and challenges that remains for the state-of-the art ASR systems. The results provide evident numeric differences between ASR systems and human perceptions, especially for younger children. Additionally, we provide suggestions for future research directions in the field.

## Palabras clave

Habla de niños; reconocimiento de voz; tecnologías del habla; WER.

## Resumen

En los últimos años, se está llevando a cabo un número creciente de estudios sobre la interacción persona-computadora, debido a las interfaces de habla generalizadas implementadas en sistemas como teléfonos celulares, asistentes personales y de automatización del hogar. Estos estudios incluyen el reconocimiento automático del habla (ASR) y la síntesis del habla, y están considerando una variedad más amplia de condiciones de las señales, como el ruido y la reverberación, y también los acentos y los efectos relacionados con la edad. Por ejemplo, uno de los desafíos clave es el desarrollo de ASR para el habla de los niños. Dado que los sistemas actuales tienen una dependencia del lenguaje y los acentos, por lo tanto, para mejorarlo, se necesitan las investigaciones de tecnologías de reconocimiento de voz adecuadas para los niños. En este trabajo evaluamos sistemas ASR comerciales para el reconocimiento del habla infantil costarricense, para usuarios con edades comprendidas entre los tres y los catorce años. Para establecer una comparación y validación numérica de los sistemas ASR para reconocer las palabras aisladas de los niños, realizamos una gran prueba de comprensión auditiva subjetiva que calcula las diferencias y desafíos que quedan para los sistemas ASR de última generación. Los resultados proporcionan diferencias numéricas evidentes entre los sistemas ASR y las percepciones humanas, especialmente para los niños más pequeños. Además, ofrecemos sugerencias para futuras direcciones de investigación en el campo.

## Introduction

During the last decade, significant progress in the field of automatic speech recognition (ASR) for general purpose devices and situations have been built and deployed, including commercial and daily-use applications.

Most of the research and implementation has been devoted to developing systems targeting adult specific speakers [1]. For children, the first studies raised attention to the poor performance of the ASR system for this population [2, 3], so increasing attention has been paid to improve this performance.

The vast majority of the research on children's speech recognition has been made for the English language, with some exceptions on other languages [4, 5] or English as a second language [6, 7].

Among the reasons for the decrease in the effectiveness of the ASR systems on children's speech, can be explained in terms of acoustic features such as higher pitch and formant frequencies, longer segmental duration, and greater temporal and spectral variability [8]. Most ASR systems based on Hidden Markov Models (HMM) or Deep Learning requires a large amount of training data, which is available recording new material that covers all the phonemes, special keywords and vast vocabularies.

These materials are more readily available or produced for adults. For that reason, general purpose ASR systems trained with specific data of adults can be affected by the spectral and temporal variability that characterize the developmental changes in speech production of children.

The pursuit of better systems is motivated by the tremendous potential in children's education, with a wide variety of possible applications ranging from pronunciation training applications to educational games [5]. For example, child-robot interaction is an area with potential contributions to domains such as health-care, education and entertainment [9]. This interaction is expected to occur in the most natural form of communication for humans that is listening, understanding and speaking.

According to [1], word recognition errors of ASR systems can be 100% worse for children's speech particularly at early childhood. The temporal and spectral characteristics that affect children's speech recognition can be aggravated by variabilities introduced by accents or regional changes in speaking styles.

For these reasons, all current ASR systems and the development of speech synthesis or other speech technologies have a dependency on language and accents. In particular, in the case of Costa Rican Spanish, there is little work published in terms of the performance of ASR systems, and even less for children's speech from this population.

In this work, we conducted an exploratory study on the performance of several commercially available ASR systems in recognizing Costa Rican children's speech. These systems can be considered state-of-the art. We also present a comparison with subjective listening tests to provide numerical differences between humans and machines. We report the recording methodology for the children and the procedures to validate the results.

### **Related work**

The effects on variability in children's speech have been studied in [5], for Portuguese children between 3 and 10 years old. The correlation of some characteristics of speech production, such as the truncation of consonant clusters, disfluencies and pronunciation quality, with ASR performance has been observed. For ages 3 to 6 years old, the recognition errors were as high as 69.9%. For children of age 6, recognition errors were found as high as 80% in some conditions [10].

For the English language, where the vast majority of studies have been made, [3] found recognition errors between 21% to 65% in ASR systems. It is clear that such error rates are unacceptable for general purpose applications or for children-computer interaction, such as the one proposed in [11], consisting

of a conversational Nao robot adapted for children's interaction. Some of the relevant features in such interactions include the ability to adapt to an individual child's language proficiency, respond contingently, both temporally and semantically, provide effective feedback and monitor children's learning progress [12]. The necessity for developing speech interfaces for more languages, which consists of ASR system, speech synthesis, natural language processing and other related technologies, has been mentioned in [13], where for most languages sufficiently large corpora of children's speech are often not available for developing speech-driven educational applications. Children would benefit from speech interfaces by using computers, tablets and cell phones, even in ages below which they have acquired reading and writing skills [14]. It is, therefore, of great interest to extend the capability of ASR systems to this speaker category.

When a proper database for developing the recognition of isolated words or sounds for children can be developed, previous studies have shown that a similar accuracy in phone recognition can be achieved with children than adults, despite the higher variability in the children's speech [15].

The rest of this paper is organized as follows: Section 2 presents the methodology we used to collect and analyze data. Also, Section 3 presents the results and comparison between human and machine word error rates, and finally Section 4 presents the conclusions.

## Methodology

This work contemplated the recording sessions with children, the transcription and the edition of those recordings. Additionally it contemplate the evaluation of the results, both with ASR and humans. The following subsections give details on these aspects:

### Speech material

For the design of the database, an interaction strategy was developed with the participating children in which they used formal and non-formal instruments of oral language assessment. The objective of this interaction was collecting isolated or two-word phrases, instead of assessing the articulation of each phoneme, which was the primary purpose of the instrument used.

The recordings contemplate words by semantic groups of high use in children's language, both in activities of daily life and within the initial school curriculum (colours, animals, food). The recordings were carried out in a professional studio.

Among the strategies used for taking voice samples when interacting with children, it is important to consider that knowledge in the area of child development is required. Not only because of the way in which it is relevant to interact with the participants but also from the aspects of perception and typical attention accorded to age in order to take full advantage of the recordings that need to be collected.

Some of the strategies employed were: the use of material with good visual contrast, utilizing game skills for word pronunciation, alternating data collection with spaces of non-formal interaction or breaks, and the use of verbal positive reinforcement.

### Speaker selection

In order to provide the inputs for the children's voice database, several recording sessions were made with children between 3 and 14 years old during this study, both male and female, and all from the central region of Costa Rica.

We split the speakers into three age groups: early childhood (ages 3 to 7), middle childhood (ages 8 to 11 years) and late childhood or early puberty (ages 12 to 14 years). We analyzed the speech recognition according to the age of the participant and also for each group. Due to the difficulties that arose in the verbal interaction with children at its earliest age, 60 words or short phrases were selected for each group.

### WER Calculation

To measure the impact on speech recognition, we implemented the Word Error Rate (WER) measure, defined as:

$$\text{WER} = \frac{(D+S+I)}{N} \quad (1)$$

where  $D$  is the number of words deleted,  $S$  the number of substitutions and  $I$  is the number of words inserted.  $N$  is the total number of words.

For ASR systems, the measure was implemented in Python, and for human listeners, the calculation was done manually.

### Automatic Speech Recognition

For the evaluation of machine ASR, we selected four state-of-the art commercial systems, which include Google Speech Recognition, IBM Watson, Happy Scribe and Cobalt Speech Cubic. We present the result without specifying which system obtained particular WER, because we pretend to assess general machine speech recognition instead of comparing commercial systems.

Each of the ASR systems is capable to personalize and boost results in several ways. For example, keywords or expressions to spot, model selection (voice commands, video transcriptions, etc.), speaker diarization and specific accent among Spanish variants.

In the first evaluation for this population, we applied each ASR with its default settings, and did not select any of the capabilities that can possibly decrease the error rate, with the exception of the Costa Rican Spanish accent when available.

The children's recordings were presented to each ASR system with two seconds of silence between each word or short phrase. This evaluation of isolated words can represent a disadvantage in terms of the language modeling that increases the ASR's capability when the context, previous and following words are detected and used to transcribe each word inside longer sentences better.

### Human evaluation

The human children's speech recognition evaluation involved 42 listeners aged between 20 and 37 years old. Each listener was a native Costa Rican Spanish speaker and lived in the same area as the children that were recorded. The listening sessions were conducted using a simple program running on a computer and a set of quality speakers.

Each participant wrote a transcription of each word by hand, for a total of 180 recordings and more than 7000 results to organize and manually process. The words were presented randomly within the recordings of each group, in the same manner that was presented to the evaluation of ASR by machines.

At the end of the session each transcription was compared to the real words that the children intend to pronounce and the results were organized according to age group.

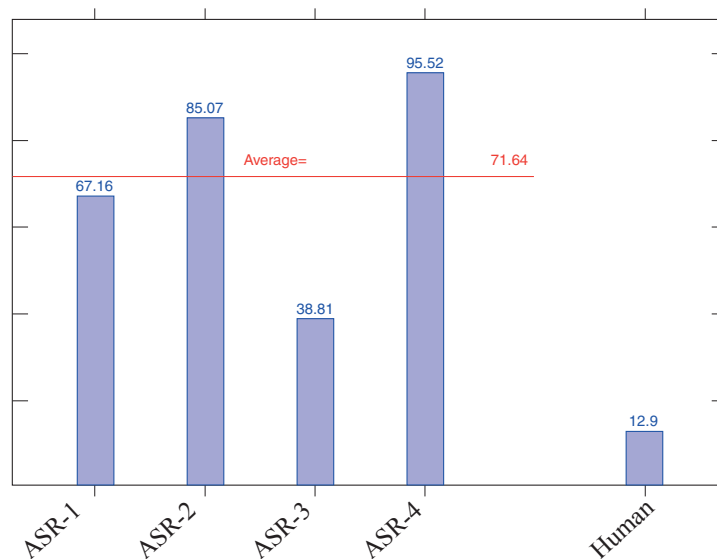
## Results and discussion

The WER results are presented and compared in Figure 1 for children between 3 to 7 years, Figure 2a for years 8 to 11 and Figure 2b for years 12 to 14. It is clear for all cases how human listeners' errors are lower than those of the ASR systems.

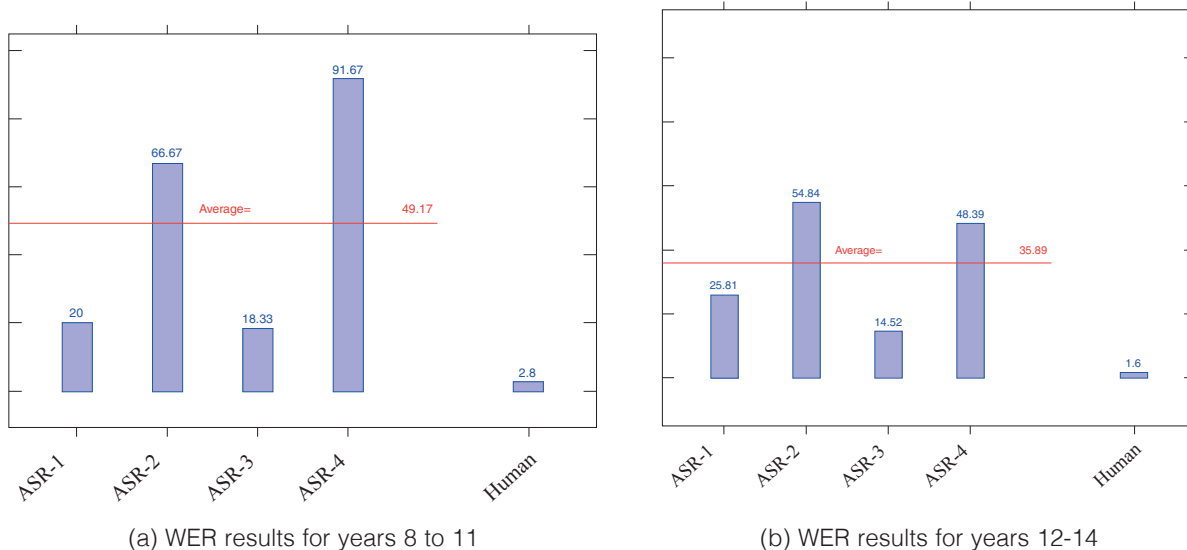
The differences increase as the age group increase. It is particularly remarkable how difficult is for computer ASR systems to recognize the words pronounced by children at the earliest age, with a range of 38.81% to 95.52%. This second percentage means that almost all of the isolated words and short phrases pronounced by children are not properly recognized. Virtual assistants, communicative robots, speech to speech translation and any other technology that relies on speech recognition can be seriously limited for this population. On the other hand, the 12.9% of WER evaluated in humans means that human listeners understood most of the words.

The differences are still high for children between the range of 8 to 11 years old. Both human and machine recognition increases its effectiveness for this second age group. With an average of 49.17%, most of the systems with automatic speech recognition will have problems understanding almost half of the messages from children in this age range, where human listeners have only 2.8% of WER. It is vital to consider also that the children's voices were recorded in the ideal, noise-controlled conditions of a professional studio.

For children aged 12 to 14, the average WER decreases to 35.89%. As expected, both for humans and machines, the WER consistently decreases as the age increases, and can be as low as 14.52% for ASR in machines in this age group.



**Figure 1.** WER results for years 3 to 7

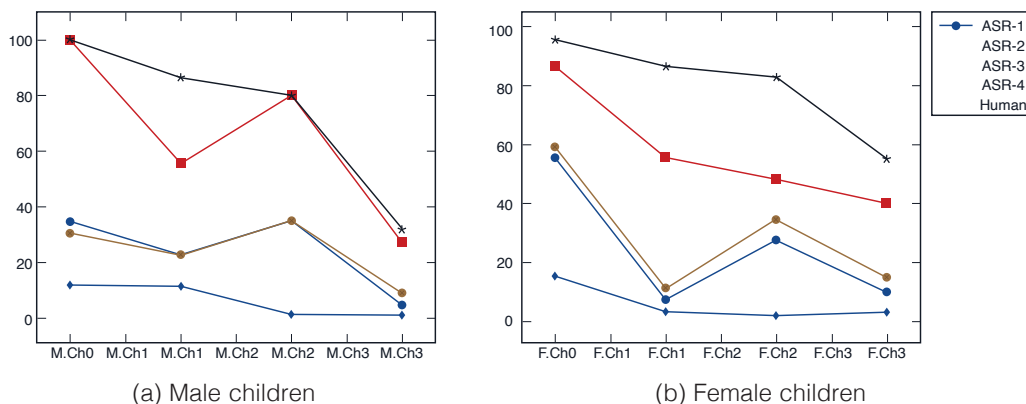


**Figure 2.** WER results for years 8 to 14

This group is also the one where the ASR systems perform more similarly than the previous groups.

In Figures 3a and 3b human and machine recognition performance is compared, according to gender. In all machine evaluation, the trend of recognition capability is the same for all systems, with similar variability for both female and male children. In general, the female Costa Rican voices present higher errors in early and late childhood when compared to their male counterparts. In comparison to the machine recognition, this trend is not present in the human listeners, where there are similar level or recognition errors at both ends of the plot, and the variance is very low for all the age range.

The errors produced by the human listeners can be explained by the oral production of the children, for which the acquisition of oral language and of course in its development process. For example, the 3 to 4 years old participants are within the stage of linguistic development and the articulation skills in terms



**Figure 3.** WER according to age and gender

of acquisition of phonemes of both are located in the first stages of development, where it is typically the processes of phonological simplification [16] where the words are simplified.

In the case of 7-year-old children the acquisition of phonemes also known as phonetic-phonological development is an average of 90 % of acquisition in the lateral articulatory modes of the “l” the vibrating modes of the “r” in syllables Fricatives “Pr” and “Fr” as well as “rr” and 80 % in the use of “ll” [17]. This could lead us to suppose that the minor participants phonological development could confuse the human participants and even more the machine recognizers.

## Conclusions

In this work we performed a comparison between humans and machines for automatic speech recognition of Costa Rican children ages 3 to 14 years old. For machine recognition we use several recognized commercial systems, and for the human assessment, we conducted a large listening test.

Results show that there is still a big difference between what machines can do in this task, in comparison with humans. Particularly for children in their early years, the machine errors can be as high as 95%, which can render a system to be of no use for children/machine interaction using voice.

As age increases, the machines tend to perform much better in understanding the words in this accent. Still, there are significant differences with human perceptions, that can be as high as 50% in more transcription errors.

This information is useful to assess those differences numerically and it can lead to new research in improving the results with modifications or adaptations of the voices in the systems.

Future work includes the development of systems trained with Costa Rican children’s voices that can be competitive or perform better for automatic recognition in this population.

## References

- [1] Gerosa, Matteo, et al. “A review of ASR technologies for children’s speech”. Proceedings of the 2nd Workshop on Child, Computer and Interaction. 2009.
- [2] Russell, Martin, Shona D’Arcy, and Lit Ping Wong. “Recognition of read and spontaneous children’s speech using two new corpora”. Eighth International Conference on Spoken Language Processing. 2004.
- [3] Li, Qun, and Martin J. Russell. “An analysis of the causes of increased error rates in children’s speech recognition”. Seventh International Conference on Spoken Language Processing. 2002.
- [4] Cossi, Piero, et al. “Comparing open source ASR toolkits on Italian children speech”. WOCCI. 2014.
- [5] Hämalainen, Annika, et al. “Correlating ASR errors with developmental changes in speech production: A study of 3-10-year-old European Portuguese children’s speech”. 2014.
- [6] Adi, Derry Pramono, Agustinus Bimo Gumelar, and Ralin Pramasuri Arta Meisa. “Interlanguage of Automatic Speech Recognition. “2019 International Seminar on Application for Technology of Information and Communication (iSemantic). IEEE, 2019.
- [7] Moussalli, Souheila, and Walcir Cardoso. “Intelligent personal assistants: can they understand and be understood by accented L2 learners?”. Computer Assisted Language Learning (2019): 1-26.
- [8] Lee, Sungbok, Alexandros Potamianos, and Shrikanth Narayanan. “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters”. The Journal of the Acoustical Society of America 105.3 (1999): 1455-1468.
- [9] Kennedy, James, et al. “Child speech recognition in human-robot interaction: evaluations and recommendations”. Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. 2017.





- [10] D'Arcy, Shona, and Martin Russell. "A comparison of human and computer recognition accuracy for children's speech". Ninth European Conference on Speech Communication and Technology. 2005.
- [11] Kruijff-Korbayov'a, Ivana, et al. "Spoken language processing in a conversational system for child-robot interaction". Third Workshop on Child, Computer and Interaction. 2012.
- [12] Vogt, Paul, et al. "Child-robot interactions for second language tutoring to preschool children". *Frontiers in human neuroscience* 11 (2017): 73.
- [13] Hämalainen, Annika, et al. "A multimodal educational game for 3-10-year-old children: collecting and automatically recognising european portuguese children's speech". *Speech and Language Technology in Education*. 2013.
- [14] Elenius, Daniel, and Mats Blomberg. "Comparing speech recognition for adults and children". *Proceedings of FONETIK 2004* (2004): 156-159.
- [15] Giuliani, Diego, and Matteo Gerosa. "Investigating recognition of children's speech". 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. *Proceedings.(ICASSP'03)*. Vol. 2. IEEE, 2003.
- [16] González, M. J. *Trastornos fonológicos. Teoría y Práctica*. Universidad de Málaga: Secretariado de publicaciones. España, 1989.
- [17] Ortiz Rubia, V. *Procesos fonológicos de simplificación*. Mendoza, Universidad del Aconcagua. Facultad de Ciencias Médicas, 2007. <http://bibliotecadigital.uda.edu.ar/229>.