Speech recognition technology in CI rehabilitation

Authors: Waldo Nogueira (1), Filiep Vanpoucke (1), Philippe Dykmans (1), Leo De Raeve (2),

Hugo Van hamme (3), Jan Roelens (3)

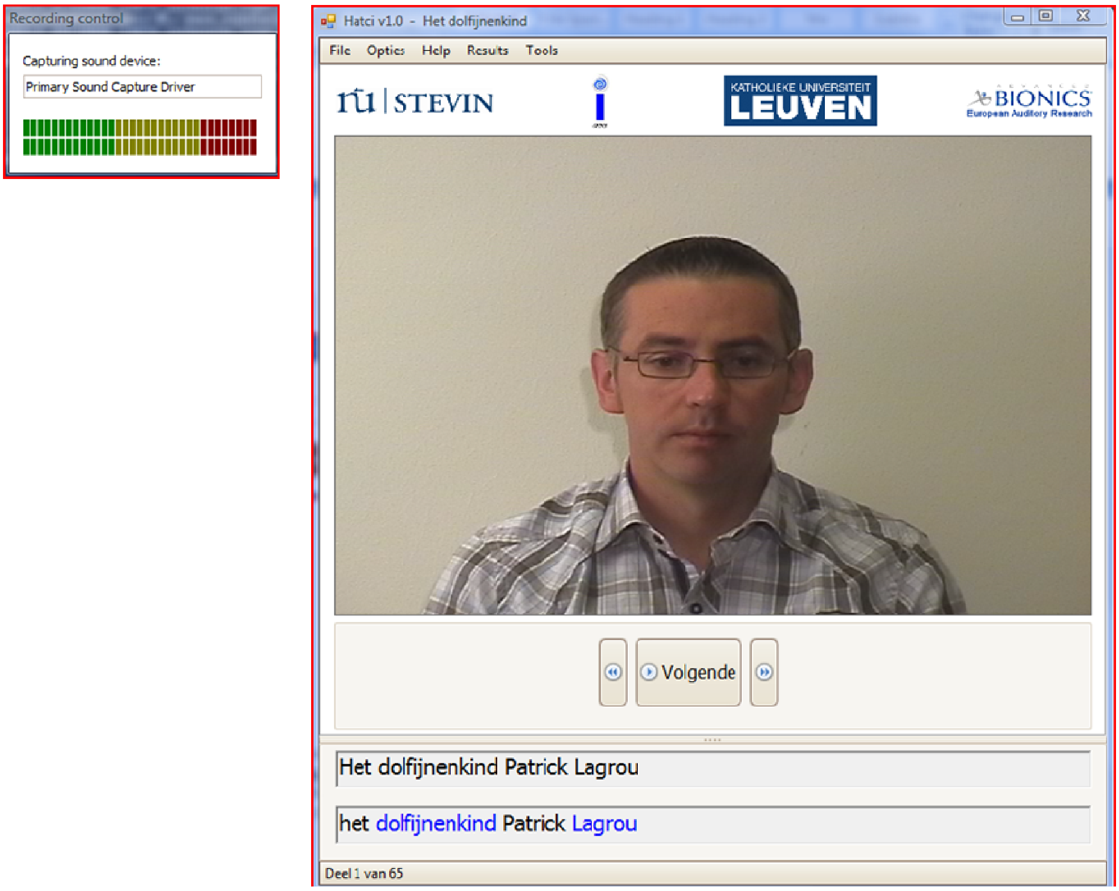
Affiliation: (1) Advanced Bionics European Auditory Research, Niel, Belgium
(2) ONICI, Zonhoven, Belgium
(3) Katholieke Universiteit Leuven, Leuven, Belgium

Introduction:

eHealth is a topic that is gaining increasing interest throughout all sectors of health care, driven by a constant drive towards greater efficiencies and quality, and by the fact that patients desire a more active, prominent role in their therapeutic care. IT and computer technologies can support this evolution in a structured way. A good example can be found in cochlear implant care. Children could benefit in the development of their auditory and language development skills by more intense exercises. However, there are relatively few speech therapists, operating in specialized centers. If children could perform exercises on the computer, in their own home or school environments, the care could be intensified in a playful manner, while at the same time keeping the cost acceptable.

Methods:

Over the lasts decades speech and language computer technologies have evolved to a mature stage with many successful products, such as automatic speech recognition software (ASR)[1]. ASR transforms audio signals into text by means of a software algorithm. In the HATCI project, a software tool that allows cochlear implant users to practice and to develop their listening and speech production skills has been designed. The HATCI is interactive and provides text, sound and video samples. The new software tool has been designed based on speech tracking [2]. First, it presents sentences playing them using an audio or/and a video stream. After each sentence, the user repeats the sentence and HATCI captures the voice. An automatic speech recognizer is used to evaluate his/her hearing and speech production abilities. Feedback about the user's abilities is given after each sentence or at the end of a list of sentences depending on the configuration selected, test or training. Figure 1 presents the main Graphic User Interface (GUI). After each session the results of the speech racking task are stored in a database. The GUI offers the possibility to create a report with all the information regarding the task, i.e. the sentences played, the correct and non-correct sentences and words repeated by the kids, the audio files containing the voice recordings, etc.



Capturing sound device:

Primary Sound Capture Driver

Het dolfijnenkind Patrick Lagrou

het dolfijnenkind Patrick Lagrou

Deel 1 van 65

Figure 1: The HATCI Graphic User Interface. On the left the voice recording component. On the right the speech tracking component. On the bottom of the right picture, the text on the top presents the original sentence uttered on the video file. The bottom text presents in black color the correct words repeated by the kid and in blue color the wrong words.

The design and development of the HATCI has been divided into three parts. First, the ONICI center has developed texts specially designed for children. Second, The ESAT institute of the Katholieke Universiteit Leuven developed and configured the Automatic Speech Recognizer for these particular texts. And third, Advanced Bionics developed a graphical interface and its interface to the ASR.

The texts were stories designed in collaboration between the ONICI center and two Flemish writers: Lagrou and Herwerkt. These tales were created using relatively easy grammatical sentences and also keeping in mind that they should be interesting for kids aging from 8 to 12 years. The tales were created in Flemish language.

For each text, a video was recorded were an actor was narrating the story. The audio contained in these video files, as well as the videos were segmented into short sentences keeping the meaning of the story.

For each sentence, a professional speech therapist predicted the errors the kid likely make. Several recordings of children were made using the HATCI speech tracker and used to validate the most common errors produced. For example, for the sentence "het hondje van een ander " the following alternatives were annotated (het/dat/de/een) hondje van een (ander/andere). In this example, the word 'het' could be substituted by the word 'dat', 'de' or 'een', and the word 'ander' could be substituted by the word 'andere'.

These common errors were used to build a grammar graph for automatic speech recognition. The Automatic Speech Recognition engine was the SPRAAK [3]. The SPRAAK was used to select which words from the grammar were understood by the kid. The grammar graph consisted of a number of states equal to the number of words of the sentence and a number of archs that connected the states (Figure 2). The amount of archs depended on the number of substitutions annotated. For each state, the original word and the deletion of the word (silence) was always implemented in the graph. In the case of 'het hondje van een andeer' the graph was composed by 5 states. Each arch was associated with a cost. This cost was different depending on the error made by the kid for each particular word. The costs are presented in Figure 2 in parenthesis together with its corresponding word. In the example shown in Figure 2, if the word uttered by the kid was correct, the cost given was 0. If the word pronounced was a substitution by another word, a penalty cost of 10 was given. A deletion (subsitutiton by silence) obtained a penalty cost of 5. Other errors as insertions, or changes in word order were also considered in the grammar graph but are not shown in the example. The automatic speech recognizer added an additional cost based on the acoustic features extracted for the voice of each kid. The task of the ASR consisted on selecting the most likely path of the graphs, i.e. the path with minimum cost. The path selected by the ASR yield to the most likely sentence uttered by the kid. This means that the ASR could only recognize the original words of each sentence and the possible errors predicted by the speech therapist. The different costs given to deletions, insertions or substitutions were the possible tuning parameters of the ASR. These parameters were optimized to maximize the performance of the ASR using several experiments with 8 kids.

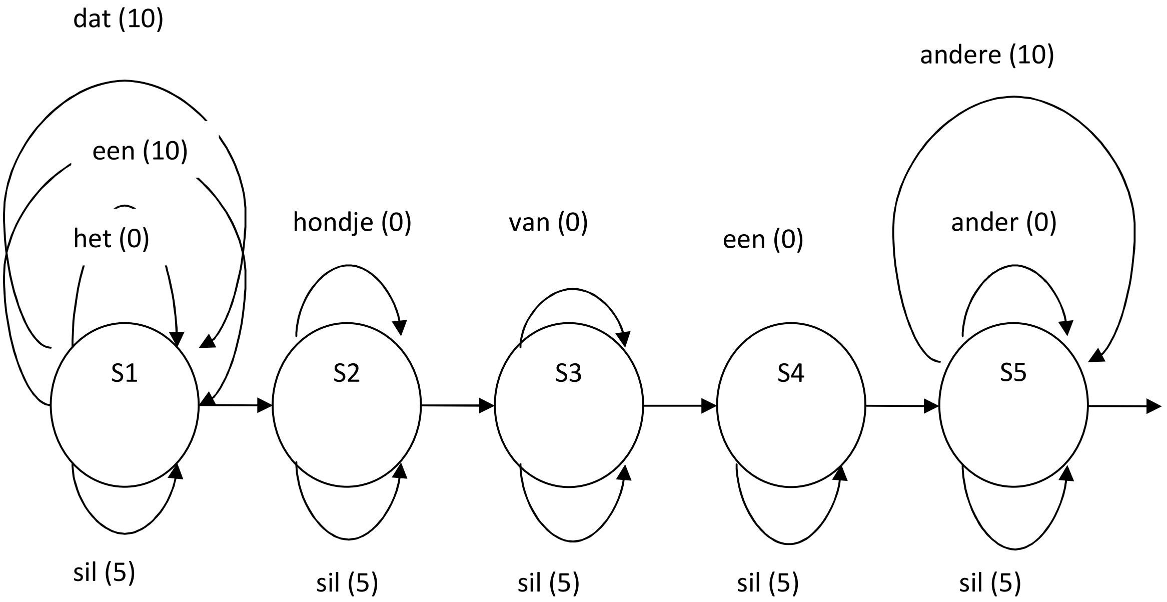


Figure 2: Grammar graph with associated costs for the original sentence 'het hondje van een ander'. The graph is composed by states and archs. The deletion of the words are represented by 'sil'. In each state different words are predicted with their associated costs in parenthesis.

Results:

The HATCI and its ASR were evaluated based on recordings of children. These recordings were annotated, i.e. the orthographic transcription was determined by an expert. These human annotations were compared to the annotations performed by the Automatic Speech Recognizer. In this report, we will first test the performance of the speech recognizer as it was used during those recording sessions. The performance of the automatic speech recognizer was evaluated with 6 kids.

Figure 3 gives the performance results of the automatic speech recognizer. Each text is represented by one dot in this graph. Every patient has his/her own color. The horizontal axis is the false alarm rate, i.e. the number of words labeled as incorrect (while they were correct), divided by the number of correct words (expressed in percent). The vertical axis is the detection rate, i.e. the number of words labeled as incorrect (while they were incorrect), divided by the number of incorrect words (in percent). An ideal system would be in the upper left corner.

The ASR engine worked reasonably well on most of the children's voices. To achieve an acceptable rate of false alarms of 5%, the detection rate (number of correctly detected errors) is ranging between 60-70%.

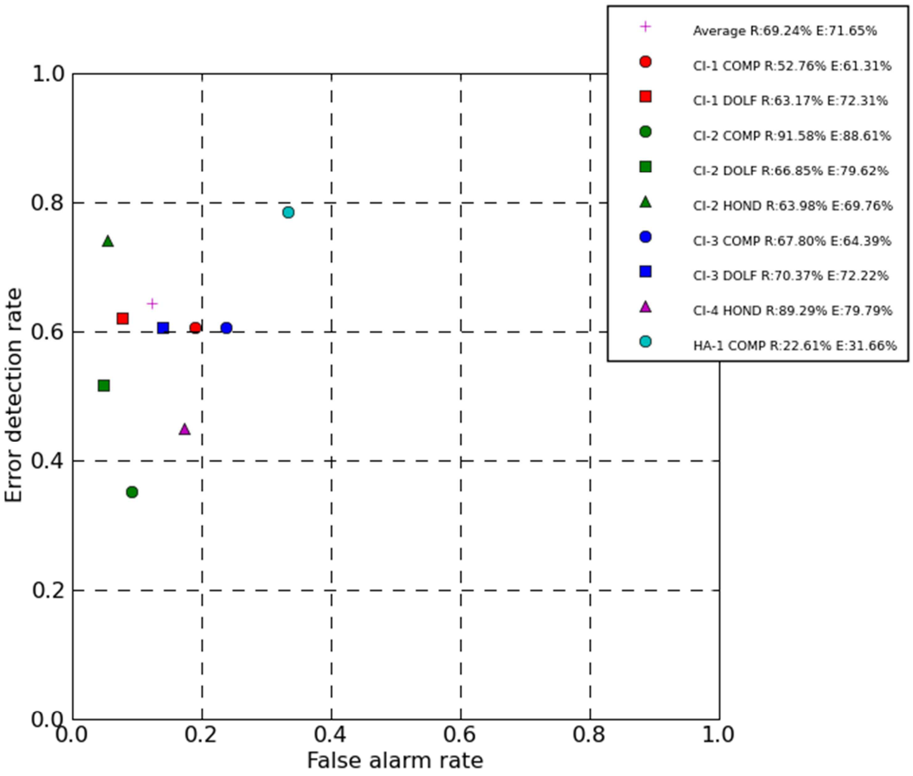


Figure 3: Performance of the Automatic Speech Recognizer used in the HATCI software tool.

Discussion:

The project has shown that ASR can certainly play a useful role in providing computer support for rehabilitation of these children. The last test sessions with the children were very encouraging. However ASR performance remains vulnerable to the quality of the incoming audio, which is not always guaranteed with a standard laptop PC. We used external microphones and audio cards, and integrated a sound level monitor in the application. These measures are possible in a school environment, but may render successful deployment in a home situation impractical.

It has to be remarked that the automatic speech recognizer was designed for the Flemish language. This means that it is not possible to use this software tool for other languages than Flemish.

Conclusion:

Speech computer technologies are a promising method to support language development and evaluation of speech therapies in children and adults.

The ASR engine used in the HATCI application worked reasonably well on most of the children's voices. To achieve an acceptable rate of false alarms of 5%, the detection rate (number of correctly detected errors) is ranging between 60-70%. Several possibilities can improve performance: penalty tuning, error prediction grammar refinements and more rigorously structured text materials.

Acknowledgement:

The HATCI project is financially supported by the Stevin program of the Dutch Language Union.

References:

[1] Huang, X, Acero, A, Hsio-Wuen, H, " Spoken Language Processing: A Guide to Theory, Algorithm, and System Development", Prentice Hall, ISBN 0-13-022616-5, 2001.

[2] DeFilippo, C. L., and Scott, B. L.: ''A method for training and evaluation of the reception of ongoing speech,'' J. Acoust. Soc. Am. 63, 1186–1192 (1978).

[3] Demuynck, K, Roelens, J, Compernolle, D.V, Wambacq, P: "SPRAAK: an open source SPeech Recognition and Automatic Annotation Kit", *http://www.spraak.org.*