SIVHA, visual speech synthesis system

Blanco Y., Cuellar M., Villanueva A., Lacunza F., Cabeza R., Marcotegui B.

Signal Theory and Communication Area.
Electrical and Electronical Department
Public University of Navarra

ABSTRACT

This paper presents SIVHA, a high quality Spanish speech synthesis system for severe disabled persons controlled by their eye movements. The system follows the eye-gaze of the patients along the screen and constructs the text with the selected words. When the user considers that the construction of the message has been finished, the synthesis of the message can be ordered. The system is divided in three modules. The first one determines the point of the screen the user is looking at, the second one is an interface to construct the sentences and the third one is the synthesis itself.

1. INTRODUCTION

In some nervous diseases the human nerve system is damaged and the movement capacity is progressively lost, including the ability to talk. These patients **don't** have any easy way to communicate with their environment. Usually they have to resort to use their eyes as the only way to communicate. The aim of SIVHA is to implement an automatic, comfortable, easy and practical system to communicate controlled solely by eye movements. The design and development of new interfaces is an expanding technology area and some new PC interfaces for disabled are being investigated. [1] But it is just the beginning and the current interface offer is very poor and most of the proposals are quite uncomfortable for the user since they need a physique connection between the patient and the PC.

2. SYSTEM ARQUITECTURE

The idea is to use the eye-gaze as the mouse of the PC to control an easy and fast speech synthesis system. To reach these goals the system must detect in real time the point in the screen the user is looking at without disturbing and using the possibilities that the interface technology brings us today, the construction of the speech must be easy and optimised in time. Finally the synthesis must be natural, intelligible and should have an appropriate voice according to the user's sex and age.

SIVHA is divided in three modules connected via TCP-IP to allow the distributed configuration of the system. The first module is called eye-tracker, the second one sentence builder and the third one is the speech synthesis.

2.1. Eyetracker

As mentioned, the system should determine the exact point of the screen the patient is looking at without disturbing. Determining this point is equivalent to track the eye movements and associate the movements to screen points. [2]

There are several techniques to track the eye movements, from the most intrusive ones, like electromagnetic measures using special contact lenses to the absolutely non intrusive methods consisting in image processing techniques. Once all these methods were studied, it was chosen to use the combination of two image processing techniques to track the eye movements, the determination of the centre of the pupil and the first reflection of purkinje. The use of these two methods assures a good estimation of the point the subject is looking at, even when the head is moved. Both methods are based in image processing and consequently both of them are non-intrusive and comfortable for the user.

The point the subject is looking at is unequivocally determined by the relative position of the centre of the pupil and the reflection in the cornea of a light source focused to the eye, the so-called first reflection of Purkinje. To calculate these two points, the centre of the pupil and the reflection in the cornea, a camera is constantly recording the subject and the acquired images will be continually processed. The camera must be situated in front of the subject and close to the screen in order to include in the image all the area of interest. To generate the reflection a light source must be chosen. This light illuminates constantly directly the eyes and must meet security and comfort criteria. To provide comfort a non-visible wavelength in the infrared band was chosen, and the camera must be able to record it. With regard to the security, the power must be limited to avoid eye damage but it must be enough to be recorded by the camera. Considering all these restriction a high power led of 8 mW was selected. The ordinary leds, as contrasted with lasers, present an excessive beam opening and collimate lens is needed to limit the reflection area in the cornea. The led is positioned in the middle of the objective of the camera. In this manner the light of the led, after entering the pupil, returns and fall directly upon the objective of the camera and all the pupil appears illuminated in the image, the so called red-eyes effect. This situation of the led doesn't affect to the image; it doesn't appear on it due to its small size and the fact that the camera is focusing the subject placed to 0.5-1 meter of the camera. Finally to avoid noises in the acquired images due to other light sources a filter of wavelengths in the visual band is needed. This allows the system to be independent of the illumination conditions. In the figure appears one of the resulting images.

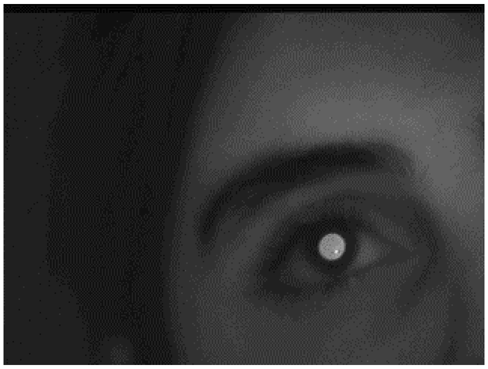

Figure1: One of the acquired eye images, where the pupil and the reflection in the cornea appear more illuminated due to the led illumination.

Once the image has been acquired they must be processed to found the two points of interest, the reflection and the centre of the pupil. The light of the led enters the eye and reflex in the cornea producing a high intensity point in the recorded image. Detecting this point is relative easy using a pixel intensity threshold. In order to find the centre of the pupil, the pupil contour must be found first. Due to the position of the led, the pupil will appear also more illuminated than the rest of the image and the grey level of its pixels will be between those of the rest of the image and the ones of the reflex of the cornea. Consequently a new threshold must be determined to decide which pixels belong to the pupil. These two thresholds are estimated at the beginning of the session using the histogram. During the session the histogram is also continually calculated to determine if the pupil appears in the image or something is hiding it, like eyelid when blinking. Once the pupil has been found the centre is determined attending to the centre of gravity of the pixels that belongs to the pupil.

Finally, when the centre of the pupil and the reflection are known, their relative position must be associated to a screen point. In other to do this, at the beginning of the session the calibration of the system takes place. This supposes to ask the user to look at special known points on the screen, and associate them with the calculated relative position of the reflection and the centre of the pupil.

2.2 Sentence Builder

This block is the responsible of the construction of the text. The requirements of this module are two basically: it must be easy to use and the construction of the speech must be optimised in time. We have added one more design condition: as far as possible the interface should learn the user's preferences.[3]

With these three goals in mind there were proposed three interfaces: Time driven interface, zone driven interface and button driven interface. The names refer to the way the elements are selected in the interface. In the first one when the user maintains the gaze in the same point of the screen for a while, the object situated at this point will change gradually its colour to advice the user that this object will be selected. The time needed to select can be changed by the user and is one of the parameters that the interface is able to learn. The second proposed interface, the zone driven interface, has special selection areas. When the user wants to select something in the screen, after looking at it, he or she must direct his or her gaze to the nearest selection area to indicate to the system that the last looked object is the one that he or she wants to select. The selection areas are located close to all eligible items in the interface to facility the selection. The third and last proposed interface is the button driven interface. In this case the patient is suppose to be able to do a small movement, not necessarily with hands. The object the user is looking at will be selected when this movement takes place. The movement doesn't need to be intensive as superficial or implanted electrodes could be used to detect this movement. This interface, probably, is the most similar to the today used mouse. Excluding the selection method the three interfaces are the same.

Interface description The sentences builder has three main areas to construct the sentence. The first one is a rotating dictionary with a vocabulary of 2000 Spanish words, where the user can select the wanted words. If the desired word doesn't appear in this dictionary there is also the possibility of constructing the word spelling it. For this purpose there is a second area with the alphabet. And finally there is an area where the most used words and sentences are kept to avoid the need of looking for the same words all along the dictionary constantly. The most used words will be automatically learned by the interface but the user has also the possibility of inserting or deleting in this area. All these three areas provide variable speed scrolling possibility, depending on the used scrolling zone. The three areas are connected in the sense that when one letter of the alphabet is selected the dictionary and the most used word area will rotate to show the words beginning with the selected letter. This process accelerates the search and selection of words and consequently the construction of the sentences. The sentence will appear at the bottom of the screen and the user can send it to synthesis it when it has been finished. Beside single sentences the interface allows to generate longer texts and to archive them to synthesis afterwards.

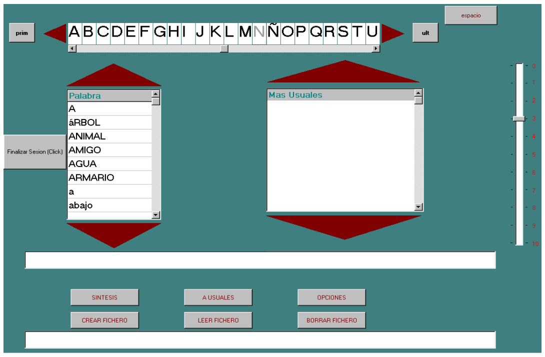

Figure 2: The sentence builder interface

Emotions and speech parameters The sentence builder offers also the possibility of inserting emotional aspects to the sentence. There is an area in the screen where the user may choose the intonation that should be used in the synthesis. There has been considered six states of mind. Beside these intonation patterns the user may change the speech mean speed, volume, fundamental frequency and the voice to be used (female, male, children, adult, old). Finally the system offers the possibility of adding emphasise to some words in the sentence.

Multi-user possibility About the learning capability, the system has been prepared to support more than one user. (f.e. it may be useful in a hospital, or for the same user with different users profiles: at work, at home…). And from one session to another it will keep the preferences of each user, such as the most used words, the speed in the time driven interface, the voice in the synthesis etc.

To achieve this task at the beginning of the session the system will ask the user to enter personal data. If the user is a new one, he or she will be introduced in the user's database and the session will begin with default parameters. But if the user has utilised previously the system, the session will begin with his or her preferences. During the session the system will actualise the preferences of the user and it will keep them until the next session.

2.3. Speech Synthesis

The last module is a Text To Speech System (TTS) and must translate the text to an acoustic wave.

To specify the requirements of this TTS, it is necessary to remember once more the aim of SIVHA: A communication facility for severe disabled persons.

Obviously, in order to communicate the system should be intelligible, but it is also necessary an appropriate intonation to understand completely the meaning of the message. Considering that the system has been designed for severe disabled persons and that SIVHA will be the permanent and probably the only way they have to communicate, the TTS should be also as natural as possible, not monotone, pleasant for the user and it should include emotional aspects as well.

Description of the TTS The TTS can be divided in three blocks: The phonetic transcription, the prosody estimation and the synthesis.

The phonetic transcription consists in the translation of the input text to a phonemes sequence. In this step the non-pronounceable symbols of the text will be eliminated and each letter will be substituted by the corresponding phoneme. This step depends on the language and in the Spanish case it consist on a small set of rules.

After the phonetic transcription, the prosody for the speech must be estimated. The prosody depends on the evolution of three parameters, volume, fundamental frequency and rhythm. The evolution of these parameters depends on the linguistic characteristics of the text and the emotional and particular aspects of the speaker. In this paragraph will be explained the prosody estimation attending to linguistics considerations. The prosody estimation due to emotional aspects will be considered in the emotions section later.

In order to assign the prosody the text will be split in smaller units, called phonic groups, according to the punctuation marks extracted in the phonetic transcription. The evolution of the three parameters will be calculated following a 2 level hierarchy. The first level refers to the average evolution of the three parameters along the phonic group. With regard to evolution of the fundamental frequency many Spanish studies affirm the existence of two inflexion points in the phonic group coinciding with the first and last strongly accentuated syllable.[4] In the used model a third inflexion point, coinciding with the second strongly accentuated syllable has been considered, providing more versatility and naturalness to the sentence. The tonic groups are classified depending on their structure and the punctuation marks. Each tonic group receives a different prosodic pattern depending on its morphology and its role in the sentence. There are several prosodic patterns categories, referred for example, to beginning, ending or continuing intentions of the sentence. Several intonation contours are available for each category: for example, ending contours can be affirmative, exclamatory, interrogative etc. Beside the intonation, the rhythm and the volume play also an important role in the prosody. The volume follows a curve along the phonic group decreasing its values at the end of the phonic group. The rhythm parameters determinate the pauses between the phonics groups and the velocity of the speech along the phonic group. The velocity depends on the large of the phonic group an inside the group it will decrease at the end.

The second level of the prosody hierarchy refers to the evolution of these three parameters inside a word. The Spanish accent involves both intonation and volume changes. The fundamental frequency will be increased progressively in the accentuated syllable and it will reach its maximum in the following syllable. The volume in the accentuated syllable will be also lightly increased. About the rhythm, the duration of the phonemes inside a word has been fixed to typical Spanish phonemes duration.

The third and last step of the TTS is the wave synthesis. Among the existing synthesis techniques, the diphone concatenative synthesis was chosen for its intelligibility and the naturalness of the produced sounds. The used concatenation methods is the TD-PSOLA (Time Domain Pitch Synchronous Overlap Add) and the used diphone database, contains 656 Spanish male voice diphones. The aim is to include more diphone database in the system to offer female and children's voices too.

Randomness The obtained speech with the prosody estimation of the two levels provides adequate results to understand the meaning and the sense of the message but it still results monotone, as the same sentences type follows the same patterns. To avoid this resulting monotony some random changes has been included in the parameters. So, in the second level the values of the evolution of the fundamental frequency, volume and duration will change around the average value in a small range, and the same word will never be said in the same way. In the first level the same theory is applied, and although the sentences follow the corresponding patterns, the standard deviation and mean values will be changed from one sentence to another. These changes maintain the attention of the listener and the resulting speech is less monotone and more natural.

Emotions SIVHA aims to be not only a communication toll but a whole expression tool for the user. So it was considered the introduction of emotions to the speech. The emotions are communicated with changes in the fundamental frequency contour, volume and rhythm, as contrast to the habitual speech.

There has been considered 6 states of mind and they are implemented increasing or decreasing the mean and standard values of the three prosody parameters from those used in the normal speech.

Control Although the system provides default parameters according to what we considered where the better ones, the user may change the mean volume, fundamental frequency and speed of the speech, as well as the voice he or she would like to use.

2.3. Connection

The three modules of SIVHA are connected via the TCP/IP. An easy protocol has been designed to allow the transmission of the data between modules.

The first module, the eyetracker determines a point in the screen. Consequently, the eye tracking system is similar to the mouse. The connection between the mouse and the computer is through the serial port of the PC. In this case the communication between the eyetracker and the PC has been designed using the TCP/IP. Before the beginning of the data transmission the IP address and the input output ports must be determined. Then, two integers corresponding with the screen point must be sent with a heading in the message known by the two modules. If X and Y are the two integers, then the message will be:

@@X#Y@@

In the same way, a protocol has been designed to communicate the second and the third module. In this case the communication is more complex, because the sentence builder most send to the TTS, the text and some parameters with the preferences of the user (voice, volume, speed, emotion…). The used message will be:

@@Parameters@@text

The heading contains the desired volume, speed and voice parameters and after the heading the message contains the text for the synthesis.

The use of the TCP/IP allows the distributed configuration of the system that could be useful in hospitals for example.

3. RESULTS

A communication and expression tool for disabled persons has been implemented. SIVHA provides a way of communicating using their eyes for those who have lost all their movement capability. The system follows the eye gaze at 20 Hz and provides a resolution of 1cm2 in the screen with a precision of 0.5cm. The system facilities the speech construction learning the preferences of the users and provides three items selection modes, in time, space and by button. And finally the speech is synthesised with TD_PSOLA technique adding an adequate prosody for the transmission of the meaning, the sense of the message, the emotions and to provide naturalness.

4. FUTURE DEVELOPMENTS

Regarding to the first module the velocity and accurateness of the eye tracking system may be improved. In the second module the interface must be tested with a large disable community to measure its efficacy and to redesign the prototype if needed. Finally in the third module, much effort must be done in the automatic estimation of prosody to provide more naturalness and a better emotion transmission.

5. REFERENCES

1. Warner D., Sale J. , Anderson T., Johanson J., *Bio-Cybernetics: A Biologically Responsive Interactive Interface* http://www.pulsar.org/febweb/papers/biocybrntks.htm

2. Jacob, R. J. K., “Eye tracking in advanced interface design” *Advanced Interface Design*, Oxford University Press, Oxford, 1995. 258-288

3. Nielsen J., “Non Command User Interface”. *Communications of the ACM.*1993,83-89

4. Garrido JM., *Modeling Spanish Intonation for Text-To-Speech Applications.* Publicacions de la Universitat Autònoma de Barcelona. 1997