Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments
Assistive Technology for All Ages
CVHI 2007, M.A. Hersh (ed.)

SUBPAL: A DEVICE FOR READING ALOUD SUBTITLES FROM TELEVISION AND CINEMA

Simon Nielsen[1], Hans Heinrich Bothe[2]

[1]Informatics and Mathematical Modelling,
Technical University of Denmark (DTU).
Tel: (+45) 26233577. Email: s060150@student.dtu.dk
[2]Centre for Applied Hearing Research (CAHR), Oersted DTU,
Technical University of Denmark (DTU).
Tel: (+45) 45253954. Fax: (+45) 45880577. Email: hhb@oersted.dtu.dk

Abstract: The primary focus of this paper is accessibility barriers for visually impaired people and people with dyslexia. Due to their disability a segment of these people have limited accessibility to the subtitle content presented on television and in the cinema. It is utterly important to be able to participate in such social and cultural events, but if the material presented is in a non familiar language they are unable to understand it. The problem primarily arises from non English speaking countries where dubbing is not facilitated such as in Scandinavian countries and the Netherlands. A solution to this problem is SubPal, a text to speech device which can be connected to the television or to a video camera. The subtitle content in the presented video stream is read aloud through a multilingual speech synthesizer. Hence the solution is applicable for television and in the cinema, in several countries. The solution comprises three major modules: The sampling of the analogue video signal into a binary image of the subtitles. The optical character recognition which converts the binary image of the subtitles into a characters that can be recognized by a computer. And finally a speech synthesizer that reads the decoded subtitles aloud. The system is quantified and a method for sampling the video signal is proposed and verified. Requirements to the optical character recognition algorithm is discussed, and parallel studies on such algorithm is referred to. The speech synthesizer is discussed in the context of user and technical requirements, and a best candidate synthesizer is evaluated. The conclusion implies potential for further studies towards a prototype.

Keywords: assistive technology, optical character recognition, sampling, speech synthesis, text to speech

1. Introduction

Approximately 6.5 percent of the western worlds population are either visually impaired or have dyslexia. As a consequence to their disability a segment of these people have difficulties reading subtitles from the television and in the cinema. This imposes a barrier when the material is presented in a non familiar language. Also people who are visually impaired to such extend that they cannot see the actual image clearly watches television. Studies, (Jønsson & Nielsen, 2006) interviewing members from the Danish organization of visually impaired people reveals that participating in such social events is of great importance. This is supported by (Tiresias, 2002), a comprehensive survey in the context of interactive digital television services for people with low vision.

As English is the dominating language (in the Western world) in television and in the cinemas, visually impaired people or people suffering from dyslexia residing in an English speaking country have better

accessibility to such material than people in non English speaking countries. Visually impaired people and people suffering from dyslexia living in non English speaking countries rely on the ability to understand English. However, learning another language being visually impaired or dyslectic requires much more effort, thus far from everybody in non English speaking countries are fluent in English.

Some of the larger European countries such as Spain, Italy, France and Germany make extensively use of dubbing to provide speech in the native language. To some extent this accommodates the needs of visually impaired people and people suffering from dyslexia in these countries. The tendency however seem to change. (Ariza, 2004) is a case study of dubbing countries with Spain as an example, concluding that subtitles are increasingly used despite the Spaniards preference for dubbing. Visually impaired people and people with dyslexia living in non English speaking countries, which does not facilitates their need by means of dubbing have very limited accessibility to the television and cinemas. Two successful technologies have emerged to accustom for this need.

In Sweden, the national Swedish television station SVT broadcasts spoken subtitles on selected programs. Subtitles are processed by a speech synthesizer and broadcasted on a separate channel. The service is made available to receivers of digital television, and requires an additional digital receiver. In Holland, FSB (Federation of Organizations for Visually Impaired People), NOS (the public broadcasting-company) and FNB (the Federation of Dutch Libraries for the Blind) have joined forces in the Spoken Subtitle Initiative (Miesenberger, Klaus & Zagler, 2002). The solution is similar to the Swedish approach: The subtitles are processed by a speech synthesizer, and broadcasted to the end user. To decode the spoken subtitles a special receiver is necessary. Commonly both solutions is that the programs made available are selected by the broadcasting station, thus limited to a segment of the programs, from a fragment of the available television channels.

SubPal is a mobile text to speech solution under development, which can be used across countries for all available television channels and in the cinema. It is a strictly end user solution, implying that the processing from video to spoken subtitles take place at the end user, thus no restrictions are imposed on the video signal.

The solution comprises a small box which can be connected to the television, or to a video camera. By connecting SubPal to a camera, and fixing the camera at a predefined distance from the screen, the same functionality is achieved as if it was connected directly to the television hence the solution is applicable in the cinema. The speech synthesizer which is used to produce the spoken subtitles, is available in 24 different languages which makes the solution generic across countries.

2. Overview

The solution is depicted in figure 1 as three main modules. A composite video signal is expected by the sampling module, which transforms the analogue video signal into a binary still image[1] of the subtitles. The binary image is fed to the optical character recognition module, where characters in the image are identified and decoded into ASCII-characters. These are reassembled to their respective word and sentences and propagated further on in the system. The last module in the system accepts an ASCII-character string, namely the decoded subtitles which are read aloud using a multilingual speech synthesizer.

Figure 1. The flow of an analogue video signal through SubPal to produce the spoken subtitles

The sampling module decodes the images of subtitles in the video signal into a binary image of the subtitles. The Optical character recognition module accepts this binary image and produces an ASCII text string of the subtitles. The speech synthesizer module accepts an ASCII text string, which is processed and read aloud. The optical character recognition module is detailed in another paper, thus

1 Formatted digitized image composed of ‘1’s and ‘0’s depicting the subtitles as binary structures.

the module is shaded. This paper reflects the initial work on the solution in figure 1. This implies the functional verification of each module, and respective methods. The interfaces for each module is appreciated and output is manually propagated as input to the next module. Thus the joint implementation of the modules is left for further studies which is elaborated on in *future work*.

3. Sampling

As depicted in figure 1, the composite video signal originating from a television or from a video camera, is accepted by the sampling module which produces a digitized binary image of the subtitles contained in the signal.

3.1 Composite video signal

The composite video signal is the most commonly used analogue video interface (Maxim, 2001). It can be extracted from the vast majority of televisions and video cameras via a standard interface such as the SCART and RCA[2,3]. There exist three television standards of the composite signal, that are used in different part of the world

- **PAL** Europe, South America, Asia, Australia and Africa
- **NTSC** North America, parts of South America and Japan
- **SECAM** France, Russia and parts of Africa

Only the PAL and NTSC standards are considered in the following as these represents the countries of primary focus. The ability to accept the SECAM standard is left for further studies. The PAL and NTSC standards primarily differ in resolution and update frequency, characteristics which are fairly uncomplicated to adhere to when sampling the signal. Thus the following analysis of the composite video signal abstracts from the variation between PAL and NTSC. The PAL signal is subject to investigation. It consists of 625 lines where approximately 575 are within the visible frame. The remaining lines are primarily used for Tele-text which also includes closed captioning (CC) subtitles. The horizontal (line) rate is 15.625 kHz, and the vertical (frame) rate is 25 Hz (Maxim, 2001).

A frame is drawn by sweeping lines across a display and retracing the scanning circuit to the next line. At the end of each frame, the scanning circuit is retraced to the first line and the process is repeated. Two methods exits for scanning lines: Interlace, and progressive. The former applies for composite video signals extracted from television and most video cameras, thus the method consider here. In interlace scanning a frame is split in two parts consisting of the even and the odd lines. A frame is drawn by first traversing the even lines and then the odd lines. A single line of a PAL composite video signal is depicted in figure 2.

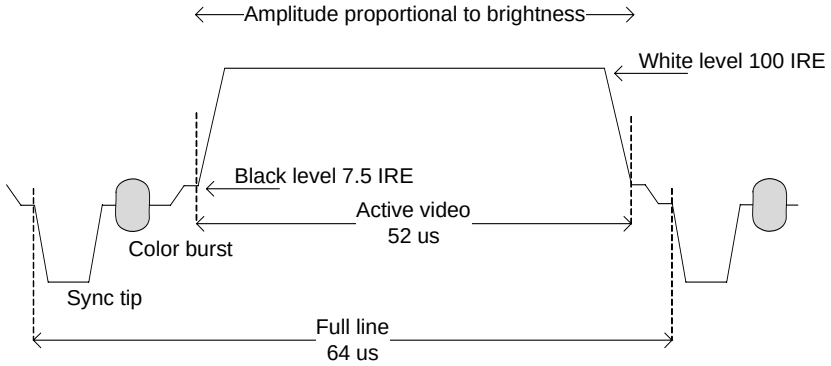

Figure 2 A single line of a Pal composite video signal.

2 (Syndicat des Constructeurs d'Appareils Radiorécepteurs et Téléviseurs) 21 pin standard interface for connecting video devices found on most European televisions.
3 Two pin connector in a single wire, typically yellow and accompanied by a white and red connector for audio.

The “sync tip” indicates a new line, causing the scanning circuit in the display to retrace and start scanning the next line. The “color burst” describes the tint (or hue) of the color, and the saturation of the color is described by the amplitude. IRE is an arbitrary unit where 140 IRE = 1 Volt peak to peak.

3.2 Method

It is interesting to note that the amplitude in the active video part is proportional to the brightness at any point on the line. Subtitles are generally represented as bright white text, often on a black background. This implies that a threshold sampling method, decoding the brightest areas of a frame would be appropriate. By extracting an Interlace Composite PAL video signal from a television and analyzing it on an Tektronix TDS1002 oscilloscope with a TDS2CMA interface, the theory is verified as subtitles in the signal are identified as high peak values, relative to the background. The hypothesis is that using two comparators[4] the subtitles can be sampled. One comparator is used for decoding the bright areas in the frame corresponding to the subtitles, and another for detecting the synchronization components of the signal. By applying a reference voltage to the first comparator just below the voltage level for subtitles, the output of the comparator should be samples of the subtitles and everything in the frame which is brighter.

To establish a measure of the necessary sampling period, the temporal spread of individual characters must be identified. By measurements on the oscilloscope the length (in time) of one line of subtitles is found to be 37 µs, with an uncertainty of ~1 µs. 38 characters (including spaces) was counted on that particular line, thus a mean length of ~1 µs per character. However the individual characters vary significantly in length. The ‘W’ is intuitively ~4 times wider that an ‘i’ consequently resulting in a 4 times greater spread. Considering a worst case scenario where all the characters in the measured line were ‘W’s, thus a ‘W’ would spread ~1 µs. Correlating this with the fact that a ‘W’ is 4 times wider that an ‘i’ imposes the necessity to detect a new letter every ~0.25 µs. Furthermore, to ensure that all characters are detected, samples are taken at least twice in each letter[5]. This results in a required sampling period of ~0.125 µs, thus a sampling frequency of approximately 8 MHz.

The hypothesis is investigated by using the oscilloscope to acquire data at its maximum sampling frequency of 25 MHz. Data is then transferred to a computer and processed with a Matlab script which emulates the comparator that detects the bright areas in the frame. To evaluate the empirically estimated sampling frequency of 8 MHz, lower sampling frequencies are simulated by removing every second sample from the data. This allows sampling frequencies of 25 MHz, 12.5 MHz, 6.25 MHz etc to be simulated. The method is illustrated in figure 3 which is a Matlab plot of one line of the acquired data. The threshold is initially selected as 0.6 Volt based on visual assessment of the signal.

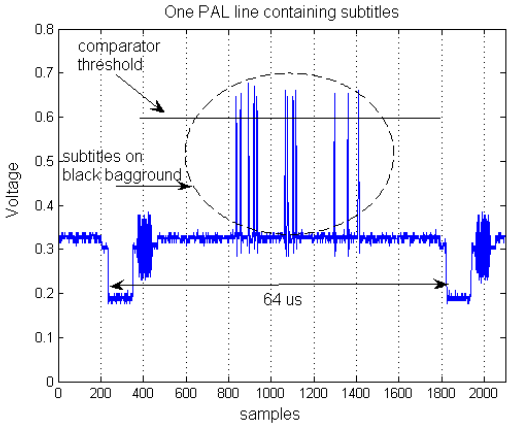

Figure 3. A single line of the PAL composite video signal

4 Electronic device that produces an output indicating which of its two inputs is greatest.
5 The alignment of characters and sample times are not synchronized, thus the worst case scenario of sampling at 0.25 µs is that samples are taken just in between characters. This can be illustrated as (‘i’|‘i’) where ‘|’ depicts the time the sample is taken. Decreasing the sampling period to just below 0.25 µs should be sufficient, however the method is a rough estimate hence sampling twice in each character should cater for uncertainties.

As shown in figure 3, data is acquired with the oscilloscope at a sampling frequency of 25 MHz, and plotted in Matlab. 1600 samples at a sampling frequency of 25 MHz, equals a line length of 64 µs, similarly the maximum peak value just below 0.7 V corresponds to 100 IRE which describes bright white in the composite signal. Thus it corresponds nicely with the theory of the composite signal depicted in figure 2. To emulate the comparators a threshold value is selected as depicted in the figure. Only values above this threshold will be extracted, thus the functionality of a comparator.

All the lines in the frame containing subtitles are traversed, and values above the threshold extracted. The spatial spread of two lines of subtitles in a frame is identified on the oscilloscope as being ~60 lines similar to the one in figure 3. As the signal is interlaced, the even and odd lines must be interleaved correctly to compose an image of the subtitles. Applying this method on the sampled data after decimating it to 12.5 MHz and 6.25 MHz yields the images depicted in figure 4a, and 4b respectively. The subtitles read: “When you say mice, do you mean those little white creatures with whiskers” .

In figure 4 the subtitles are unclear due to resizing of the image. In the real image the individual characters and words are perfectly discriminable. Figure 4a reveals that sampling at 12.5 MHz is sufficient as the subtitles are easily identified, but sampling at 6.25 MHz which is illustrated in figure 4b, is not sufficient as the subtitles can not be identified clearly. Thus the calculated sampling frequency of 8 MHz is a realistic estimate. This concludes that the hypothesized sampling method can be applied to decode the subtitles from a composite video to images of the subtitles presented as binary structures. The method will be implemented on an FPGA comprising the interface depicted in figure 1. Thus accepting a composite video signal, producing binary images of the subtitles that the OCR algorithm accepts.

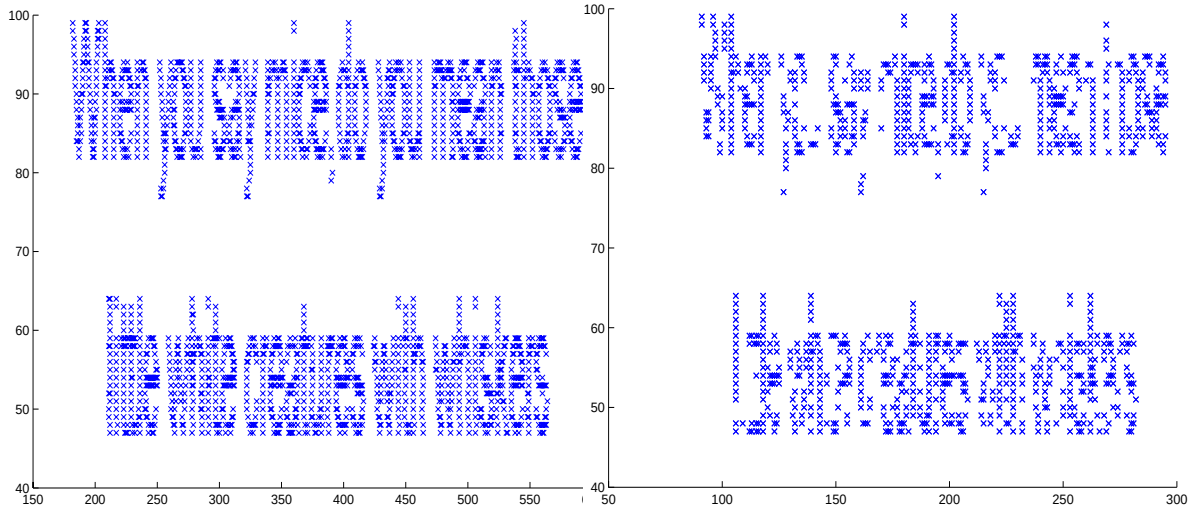

Figure 4. The method depicted in figure 3 is applied to all the lines in a frame which contains subtitles, and the output is composed to an image of the subtitles.
a). Method applied to data sampled at 12.5 MHz . The text is clearly identified which implies that the sampling frequency is sufficient.

b). Method applied to data sampled at 6.25 MHz. The text cannot be identified, thus the samplings frequency is to low.

4. Optical Character Recognition

The OCR module accepts a binary image of the subtitles and produces an ASCII string which is expected by the speech synthesizer. The OCR is not detailed in this paper, but (Jønsson & Bothe, 2007) describes the development of an OCR which is targeted for SubPal and thus comprises with the interface in figure 1.

5. Speech Synthesizer

Commercial synthesizers have been surveyed with respect to two categories: User requirement, and Technical requirements. User requirements have been gathered from visually impaired people in Denmark. The majority of their text to speech products features the Danish speech synthesizer from "Mikrovaerkstedet"[6]. The generally opinion is that the quality of speech is to bad (Jønsson & Nielsen, 2006). The synthesizer is a concatenative synthesis of the type Diphone. One of the major problems in concatenative synthesis is the audible discontinuities between successive units. The Diphone synthesis generally only contains one instance of each diphone in the database (Klabbers, van Santen, & Kain, 2007). Thus these factors compromises the speech quality. The Unit selection synthesis is better alternative with respect to speech quality. It is also a concatenative synthesis but utilizes a large database of speech units, thus producing a more natural sounding speech (Clark, Richmond & King, 2007). The selection and concatenation of speech units from a large database consequently compromises the performance, which is a critical factor as the system must cater for a worst case scenario where a new set of subtitles arrive with approximately 1 second intervals. The most promising candidate for the solution is the unit selection speech synthesizer from Acapela[7]. It is available in 24 different languages and supported for various platforms. Further, It provide nice tradeoff options between performance and speech quality which is essential to meet the real time requirements imposed by the continuous stream of subtitles.

The speech synthesizer has been evaluated on a PC with a special developed application that accepts an ASCII text strings from a Serial connection. The quality of speech is extraordinary compared to "Carsten". It is very versatile across different languages, the process of changing language merely involves loading another voice. Thus from a user perspective the speech synthesizer from Acapela is very feasible. From a technical perspective further studies is necessary to cover response time measurements relative to compromising the quality. However promising results have been gathered from the Melfo project[8] an embedded text to speech solution where the synthesizer from Acapela is used. They have utilized its flexibility to achieve a good quality and performance.

6. Summary and Future Work

All of the modules (figure 1) needs to be implemented to a combined solution. An embedded onboard device comprising a microprocessor and special purpose hardware is found suitable. The sampling method is verified and implementation should be trivial. Targeting this for an FPGA is appropriate with respect to the high sampling frequency. The optical character recognition algorithm detailed in (Jønsson & Bothe, 2007) is similarly intentioned to be implemented in hardware, which facilitates parallelism and thus greater performance. The speech synthesizer is supported for at number of microprocessor platforms. Characteristics of mobile devices such as low power consumption, and small physical dimensions are facilitated by the Intel XScale processor[9], thus making it a promising candidate. Once a suitable microprocessor is found, optimizations with respect to the performance - quality tradeoff must be conducted.

The studies reveal potential for a commercial solution, however considerations must be made with respect to the transformation from analogue to digital television. In Denmark the inherent digital broadcasting net is activated from 2009. The digitalization of the signal eventually implies a shift from "burned in" and "tele-text" subtitles to DVB (Digital Video Broadcasting) subtitles. Burned in subtitles are superimposed on the video signal prior to broadcasting and subtitles primarily used in non English speaking countries, for instance for viewing films in English. Tele-text subtitles are broadcast with the tele-text and therefore can only be received if the viewer has a teletext decoder. These subtitles are mainly used in English speaking countries where subtitles are provided for hearing impaired people. This makes the decoding of subtitles from the television fairly easier, but requires a different approach. The DVB subtitles are merely broadcasted as coded character strings. Thus the optical character recognition module is not required, but instead a decoder module comprising with the format is necessary. However the approach proposed in this paper is considered generic as the composite

6 Mikro Vaerkstedet, "The speech synthesizer, Carsten" from 2002 http://www.mikrov.dk/sw7986.asp
7 Acapela Group HQ speech synthesizer. Demonstrations available at: http://www.acapela-group.com/demos/samplesHQ.asp
8 Melfo project. Initiative on assistive technology for visually impaired people and people suffering from dyslexia from 2006: http://www.crossroadscopenhagen.com/Nyheder/Nyhedsarkiv/Gennembrud_i_MELFO
9 Intel XScale processor. Special designed for embedded devices: http://www.intel.com/design/intelxscale/

video signal extracted from a television, regardless of it being analogue or digital, will contain subtitles that can be decoded with presented method.

References

Ariza, M.C.G. (2004). A case study: Spain as a dubbing country, *Translation Journal*, Volume 8(3).

Clark, R.A.J., K. Richmond, S. King, (2007). Multisyn: open-domain unit selection for the Festival speech synthesis system, *Speech Communication*, vol. 49, pp. 317-330.

Jønsson, M. and H.H. Bothe (2007). OCR for detection of subtitles in television and cinema, *CVHI*.

Jønsson, M. and S. Nielsen (2006). Interviews and conclusions from meetings with chairpeople from departments of accessibility, with courtesy to “Michael Jensen” from the Society of Visual Impaired, www.dkblind.dk, Denmark.

Klabbers, E., J. P. H. van Santen, and A. Kain (2007). The Contribution of Various Sources of Spectral Mismatch to Audible Discontinuities In a Diphone Database, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15(3), March 2007

Maxim (2001). Dallas Semiconductor, Video basics: http://www.maxim-ic.com/appnotes.cfm/appnote_number/734

Miesenberger, K., J. Klaus and W. Zagler (Eds.) (2002) *ICCHP*, LNCS 2398, pp. 295–302, Springer-Verlag Berlin Heidelberg.

Tiresias (2002). Visual disability survey on “Interactive Digital Television Services for People with Low Vision”: http://www.tiresias.org.