International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:1, No:2, 2007

A Smart-Visio Microphone for Audio-Visual Speech Recognition "Vmike"

Y. Ni, and K. Sebri

Abstract**—**The practical implementation of audio-video coupled speech recognition systems is mainly limited by the hardware complexity to integrate two radically different information capturing devices with good temporal synchronisation. In this paper, we propose a solution based on a smart CMOS image sensor in order to simplify the hardware integration difficulties. By using on-chip image processing, this smart sensor can calculate in real time the X/Y projections of the captured image. This on-chip projection reduces considerably the volume of the output data. This data-volume reduction permits a transmission of the condensed visual information via the same audio channel by using a stereophonic input available on most of the standard computation devices such as PC, PDA and mobile phones. A prototype called VMIKE (Visio-Microphone) has been designed and realised by using standard 0.35um CMOS technology. A preliminary experiment gives encouraged results. Its efficiency will be further investigated in a large variety of applications such as biometrics, speech recognition in noisy environments, and vocal control for military or disabled persons, etc.

Keywords**—**Audio-Visual Speech recognition, CMOS Smart sensor, On-Chip image processing.

I. Introduction

AUDIO-VISUAL coupled speech recognition using both voice and video information (talking face) is a natural extension of the pure voice based speech recognition. It can potentially improve the recognition accuracy in some difficult situations such as noisy environments. Visual information is thought to be actively used in human speech recognition activity [1]-[4]. For example a person in a very noisy subway station is able to recognise correctly the speech of other people by observing his lips movements because the visual information on lips motion is not corrupted by the ambient acoustic noise [4]. The recent studies [1] have demonstrated that the audio-visual speech recognition systems (AVSR) can effectively improvement the recognition rate. However, two major problems have to be taken into consideration: 1) algorithms which can extract lips information in real-time and 2) hardware integration of the audio and visual captures with good temporal synchronisation.

In contrast to recent research in video-audio coupled speech recognition studies, we will not use sophisticated image analysis oriented extraction of the lips shape and motion features. We think that a simple projection on the axis X & Y of the lips region could give a salient information to increase speech recognition accuracy in hostile environments [5], to enhance speech analysis based biometrics against simple playback attack [6][7] and to improve the reliability of speech controlled command interface in some difficult situations [8]. From our observation, when an image sensor is mounted on a microphone, a user has a natural reflex to point the sensor to the mouth with pretty good alignment. This fact simplifies considerably the problem because the lips region detection and segmentation is no longer necessary in this configuration. A typical image captured by such a device is shown in Fig. 1; the correlation between the lips opening and the projections is clearly visible.

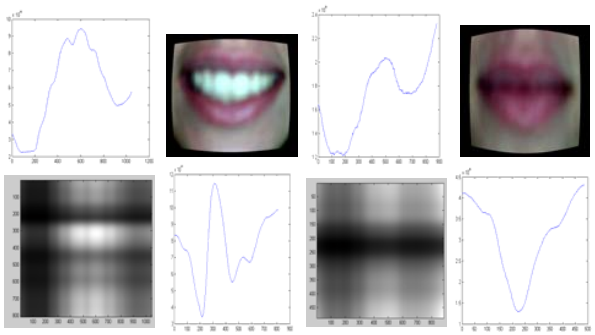


Fig. 1 Example: the lips shape can be deducted by using X/Y projections of the mouth

By this algorithm simplification, it is possible to design and implement this projection computation inside a smart CMOS image sensor. The output data volume of this smart sensor is largely reduced and comparable to that of voice signal. In this way, the video and audio signals can be injected into a general purpose computing device such as a PC via the standard stereophonic input. This device called Vmike (Visio-Microphone) will give not only natural synchronisation between image and voice but also a very simple and easy way to integrate Vmike in real applications. In the following

Manuscript received December 19, 2006. This work was supported by the Department of Electronics and Physics in the National Institute of Telecommunications and ENST Paris.

Yang Ni is with the Department of Electronics and Physics, The National Institute of Telecommunications, 9 Charles Fourier, 91011 Evry, France (phone: 0033 (0) 1 60764648; fax: 0033 (0) 1 60764284; e-mail: yang.ni@int-evry.fr).

Khaoula Sebri is also with the Department of Electronics and Physics, The National Institute of Telecommunications, 9 Charles Fourier, 91011 Evry, France (phone: 0033 (0) 1 60764661; fax: 0033 (0) 1 60764284; e-mail: khaoula.sebri@int-evry.fr).

sections, we will present the detailed implementation of this Vmike smart CMOS image sensor and give some first hand experimentation results.

II. Design and Realization of VMIKE

A. Vmike Smart Sensor Structure and Design

The Vmike image sensor is made of a photodiodes array in which the vertical and horizontal wires connect alternatively the photodiodes to X and Y projection nodes as shown in Fig. 2. This will permit not only to reduce enormously the surface of the pixel because there is no need to use buffer amplifiers to amplify signal issued from pixel array of the sensor but also it will guarantee the facility of wiring inside the chip.

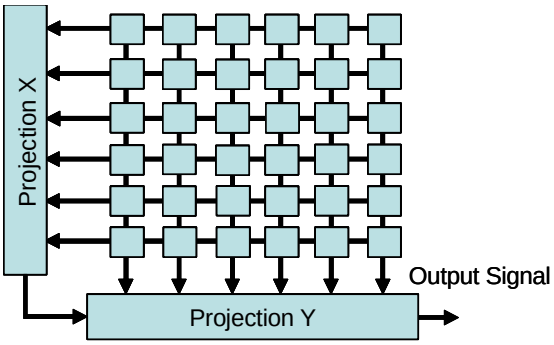


Fig. 2 Vmike smart sensor's architecture

The X/Y projections are formed directly by wiring photodiodes into the respective projection nodes. The projections are read out sequentially by using a capacitive circuit.

A differential ambient light suppression mechanism is implemented in Vmike sensor. For each image capturing, two frames are captured, one with LED on and the other with LED off. The projections of these two frames are stored at two capacitor arrays (C1 & C2) as shown in Fig. 3. The stored images will be subtracted from each other during the image readout phase by using on-chip differential amplifier.

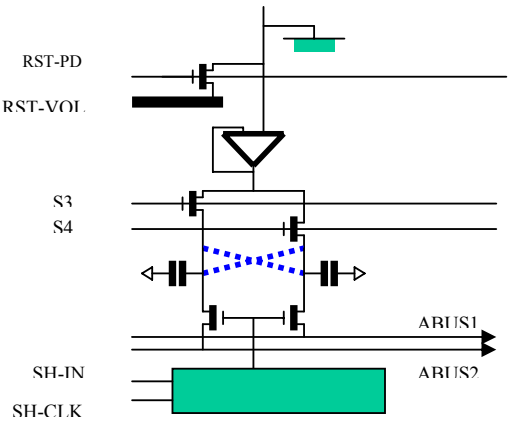


Fig. 3 Schematic of pixel unit and readout circuit

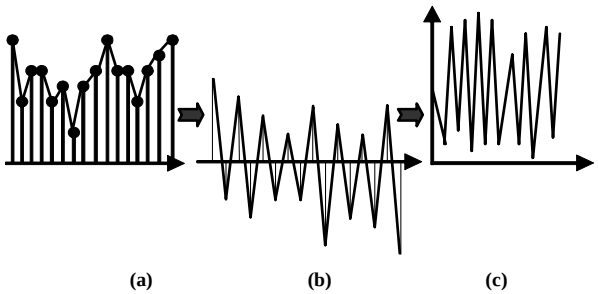


Fig. 4 (a) Original image projection, (b) Modulation by successive alternation, and (c) Modulated Signal

This differential capture is further combined with the necessary modulation of the X/Y projections signals. Because all the audio interfaces use AC coupled structure, so the X/Y projections can not be injected directly inside. Otherwise the DC component will be lost and result in waveform distortion. A modulation is necessary. This modulation is combined with the image differential readout by permutating alternatively the connection between the C1 & C2 and the readout buses as shown in Fig. 4 and Fig. 5. The carrier central frequency will be the half of the Vmike sensor's readout frequency which can be adjusted by the microcontroller generating the control signals.

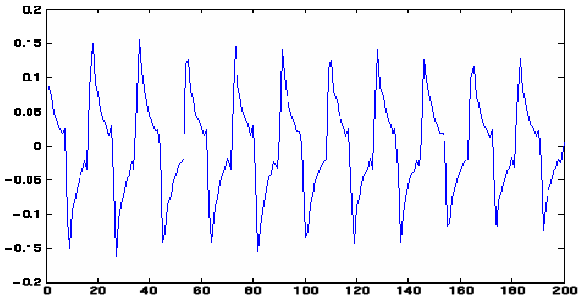


Fig. 5 Output signal of Vmike smart sensor through an AC coupling circuit

Recovery of the modulated signal may be made by direct detection from the absolute value of the modulated signal as shown in Fig. 6. The frame synchronization is detected by using the silence time corresponding to Vmike sensor's exposure time.

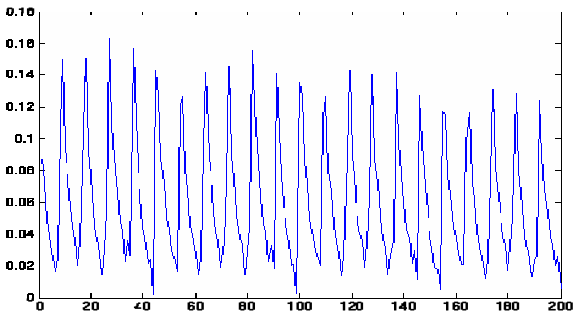
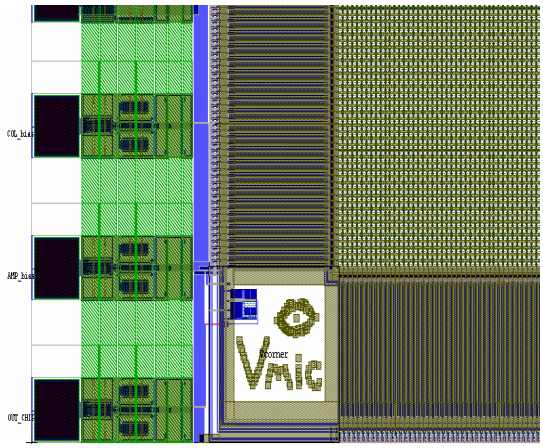
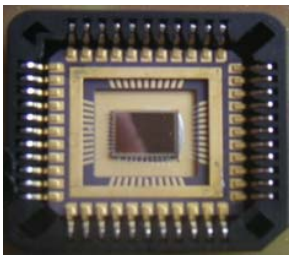


Fig. 6 Demodulated signal

We have designed and realized a 200x200-pixel prototype circuit in 0.35 µm CMOS technology from AMS via French CMP service. Pixel pitch is of about 10µm. Tanner software and Hspice are used to design and simulate this circuit respectively (see Fig. 7 (a)). The image sensor circuit was encapsulated using a traditional image sensor package CLCC. Total power consumption is less than 3mW from a 3.3-V supply (see Fig. 7 (b))



(a)



(b)

Fig. 7 (a) Layout of Vmike sensor under L-Edit editor, and (b) Vmike sensor chip

B. VMIKE Prototype

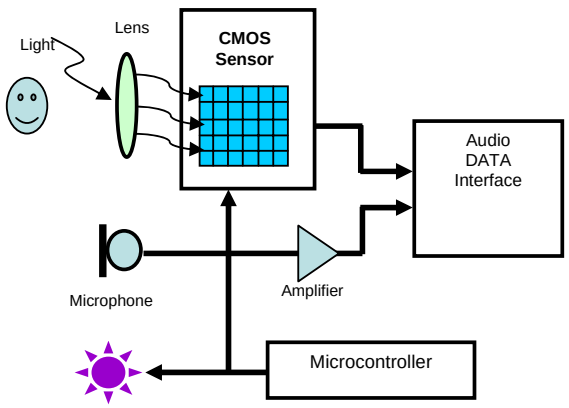


Fig. 8 Global architecture of Vmike

A prototype Vmike has been realized. The control signals are generated by a microcontroller, AT89C2051 as shown in Fig. 8. All is in form of a hand held microphone. When we speak in the microphone and the smart sensor will point instinctually to mouth region. The audio and video signals are combined by using stereophonic jacks. Fig. 9 shows the functional Vmike prototype. In this prototype the frame rate is fixed at 12 images per second, which gives a projection central frequency of 2.4 KHz. It will be increased further. The main reason to use this low speed is the fear of possible dysfunction of sound card with the fast transition in the Vmike sensor's output.

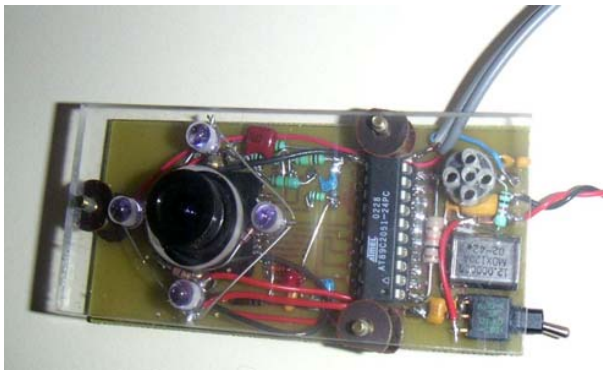


Fig. 9 Vmike Prototype

C. Preliminary Experiment

The first experiment with Vmike has been done on a PC by using an audio waveform capturing shareware, Audacity. The X/Y projection signal is stored in a file and then the demodulation has been done by using MatLab (see Fig. 10). A real-time demodulation and displaying interface will be programmed soon.

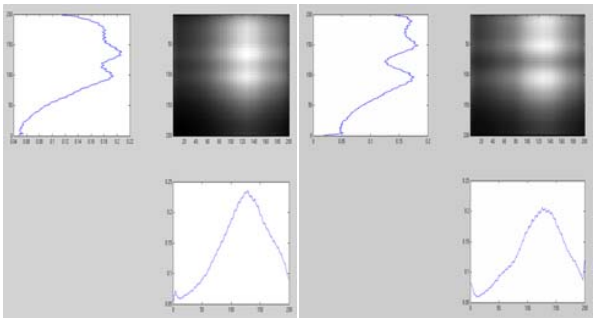


Fig. 10 Preliminary experimentation: Examples of lips image extraction

III. Conclusion

This paper presents an X/Y projection calculating smart image sensor augmented microphone Vmike which is dedicated to simplify the hardware implementation of audio-video coupled speech recognition applications. Thanks to on-

chip X/Y projection computation, the video signal bandwidth has been successfully reduced to the same as of the audio signal. A smart sensor of 200x200 pixels has been designed and fabricated in 0.35um CMOS process. A functional prototype has been realized. The first experimentation gives very encouraging results and further investigations are on the way. This small, simple, low power and low cost device can be used in many applications based on speech recognition such as biometrics, voice commands, etc.

REFERENCES

[1] Yashwanth H, Harish Mahendrakar and Suman David, "Automatic Speech Recognition using Audio Visual Cues", IEEE India Annual Conference 2004, INDICON 2004.
[2] L. Liang, X. Liu, Y. Zhao, X. Pi, and A. V. Nefian, "Speaker Independent AUDIO-VISUAL Continuous Speech Recognition", In IEEE International Conference on Acoustics, 2002.
[3] J. Huang, G. Potamianos, and C. Neti, "Improving Audio Visual Speech Recognition with an infrared Headset", AVSP 2003 - International Conference on Audio-Visual Processing, St. Jorioz, France, September 4-7, 2003.
[4] M. Kaynak, Q. Zhi, A. D. Cheok, K. Sengupta, Z. Jian, and K. Chi Chung, "Analysis of lip Geometric Features for Audio-Visual Speech Recognition", IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, VOL.34, NO.4, July 2004.
[5] C. Neti, G. Potamianos, J. Luettin, I. Matthews, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-Visual Speech Recognition", WORKSHOP 2000 Final report, October 12, 2000.
[6] www.globalsecurity.org/security/systems/voice.htm.
[7] A. K. Jain, A. Ross and S. Prabhakar, "An introduction to Biometric Recognition", IEEE Transactions on Circuits and Systems for video Technology, Special issue on Image-and-video-Based Biometrics, VOL. 14, No. 1, January 2004.
[8] www.thalesgroup.com/avionics/markets/military_aircraft

Yang Ni received the M.Sc. degree in electronic engineering from South-East University, Nanjing, China, and University Paris Sud, France, and the PhD. degree, also in electronic engineering, from University Paris Sud. He is currently a Professor in the National Institute of Telecommunications Evry, France, and his research interests include analog/digital integrated circuits for artificial vision.

Khaoula Sebri received the M.Sc. degree in electronic engineering from National School of Engineering of Sfax, ENIS, Tunisia. She is currently working toward Ph.D. degree on Microelectronics at the University Pierre & Marie Curie (PARIS VI) and The National Institute of Telecommunications, Evry, France. Her research interests include analog sensor for artificial vision and their interface Design.