Lip Tracking Method for the System of Audio-Visual Polish Speech Recognition

Mariusz Kubanek, Janusz Bobulski, and Lukasz Adrjanowicz

Czestochowa University of Technology
Institute of Computer and Information Science
Dabrowskiego Street 73, 42-200 Czestochowa, Poland
{mariusz.kubanek,januszb,lukasz.adrjanowicz}@icis.pcz.pl

Abstract. This paper proposes a method of tracking the lips in the system of audio-visual speech recognition. Presented methods consists of a face detector, face tracker, lip detector, lip tracker, and word classifier. In speech recognition systems, the audio signal is exposed to a large amount of acoustic noise, therefor scientists are looking for ways to reduce audio interference on recognition results. Visual speech is one of the sources that is not perturbed by the acoustic environment and noise. To analyze the video speech one has to develop a method of lip tracking. This work presents a method for automatic detection of the outer edges of the lips, which was used to identify individual words in audio-visual speech recognition. Additionally the paper also shows how to use video speech to divide the audio signal into phonemes.

Keywords: lip reading, visual speech, audio visual speech recognition.

1 Introduction

Automatic speech recognition (ASR) is widely used as an effective interface in many devices: personal computers, robots, mobile phones and car navigation. For ASR systems, in low noise-level environments, the word correct rate (WCR) for audio channel only is over 95%. However in noisy environments, the WCR is significantly reduced [1]. To overcome this problem, we consider a lip reading method. Automatic recognition of audio-visual speech provokes a new and challenging tasks of comparison and competition with automatic recognition of the audio speech. It's well known that the visual modality contains some complementary information to the audio modality. The use of visual features in audio-visual speech recognition (AVSR) is motivated by the bimodal nature of the speech formation and the ability of humans to better distinguish spoken sounds when both audio and video are available [2,3,4]. Audio-visual speech recognition examines two separate streams of information, in comparison to only audio speech. The combination of these streams should provide better performance in contrast with modern approaches that utilise each source separately. The issue of video characteristics extraction and fusion of audio and video characteristics are difficult

L. Rutkowski et al. (Eds.): ICAISC 2012, Part I, LNCS 7267, pp. 535–542, 2012.

problems, that generate a lot of research in the scientific community. Since E. Petajan demonstrated in his work [5] that audio-visual systems are more effective than either speech or vision systems alone, many researchers started to investigate audio-visual recognition systems [6].

The paper presents a method for automatic detection of the outer edges of the lips. In addition, it shows how to improve the distinguish ability of video sounds, by analyzing the position of tongue and how to use video speech to divide the audio signal into phonemes. Because it is difficult to evaluate the effectiveness of lip-tracking, as the tracking accuracy may be verified only by observation. Therefore performance tests where based on isolated words recognition of audio-visual Polish speech. Moreover this work focuses on hidden Markov models (HMM) and presents a method of automatic lip tracking.

2 Face, Eye and Area of the Lip Detection

The first step in the process of creating a video observation vectors of speech, is the location of the user's face in a video. Because the system was designed to operate with only one user at any given time, it is assumed that the frame contains only a single face of one individual. The process of face localization involves the reduction of entire frame to the area containing only the face. This work uses haar-like features [7,8] as the detection method to locate the face area.

After determining the coordinates of vertices rectangular mask, a new video sequence of statements containing the limited area of the image to the user's face is creates from the original sequence of frames. For a lip-reading system, it is essential to track the lip region of the speaker. This can be achieved by tracking the lip-corners. It is difficult to locate or track lip-corners alone. In order to find the lip-corners within a face, it is possible to search other facial features using certain constraints and heuristics. Some facial features are easier to locate than lip-corners. For example, within a face, the pupils are two dark regions that satisfy certain geometric constraints, such as: position inside the face, symmetry according to the facial symmetric axis and minimum and maximum width between each other.

When designating mouth image area it is well known that individual frames contains the entire face of the user. Therefore eye coordinates can be used to determine the exact position of the mouth. For this reason Gradient Method and Integral Projection (GMIP) [9] is applied to find horizontal and vertical lines of eyes.

3 Lip Edge Detection and Video Encoding Speech

During natural speech, lips move vertically, the lip corners are in place or, alternatively, move horizontally, and the distance between the inner and outer edge of the lower and upper lip remains unchanged. Therefore, the system uses only

the corners and outer edges of the lips as the main features in the process of lip tracking [10].

An important element of the extraction characteristics of the lips is to locate the lip corners. The exact position of the mouth corners is crucial, so that later on the basis of their location the outer edges of the mouth could be detection, as well as the landmark distribution. This paper proposes a method based on the specifics of lips color and shape. In this method, the localization process of lip corners is realized on a color image. Lips have a very distinct color and by properly manipulating the various components of the RGB color space, isolated borders between the lips and the rest of the face might be obtained by thresholding. Operations performed on the RGB channels could be described by the following relationship:

$$lipregion = \begin{cases} \frac{B}{G} - 1 < T1 \\ \text{and} \\ \frac{R}{G} - \frac{B}{G} < T2 \\ \text{and} \\ \frac{R}{G} - 1 < T3 \end{cases} \tag{1}$$

where: $T1$, $T2$ and $T3$ are empirically chosen thresholding values.

In this way, the values of the pixels corresponding to the specific color of the lips are set. Knowing the structure shape of lips, corners of the mouth may be designated as the extreme levels of specific color of mouth pixels. Searching for the pixels has to be done within a limited area of the mouth, near the horizontal axis of the lips.

Correct determination of the lip corners is so important that a round grid is built on its basis to determine the points on the outer edges of the lips. Based on the corners of the mouth the center of the circle can be determined at half of the distance between the corners of the lips. When moving around the circle the radius is determined as the angle α. Starting from one of designated corner we can differentiate between $2\pi/\alpha$ of rays. Then, moving along each of the rays toward the center of the circle we can designated a characteristic point of the outer edge of the lips, as the first encountered mouth pixel. The offset of rays is chosen accordingly to the accuracy with which the outer edges of the lips should be reproduce. The study assumed that each of the rays is at about 22.5 degree, which gives 16 points, including two characteristic corners of the mouth.

The system is based on HMM. For the HMM model, the input signal has to be introduced as a vector of observations, so for each frame based on the coordinates of characteristic points a symbol could be assigned that best describes the characteristics of that frame. The proposed method for encoding frames incorporates a simplified method that uses the location of each of the characteristic points of the straight line defined by the corners of the mouth. For each frame, we calculate the sum of relative distances m from all points of a straight line, defined by the corners of the mouth. We adopted 16 characteristic points, so each of the calculated relationships could be divided by 16. The sum of obtained

value, when multiplied by 100 is in the range from 11 to 60, obviously if properly located in characteristic points on the outer edges of the lips:

$$y = \frac{\sum_{i=1}^{N} \frac{m_i}{d}}{N} \cdot 100 \tag{2}$$

where: N - is the number of points, m and d - see Fig. 1.

Fig. 1 shows scheme of assumed location of corners of mouth, definition of external edges of mouth and video speech encoding method.

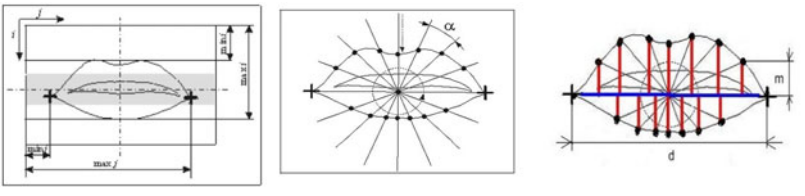

Fig. 1. Scheme of assumed location of corners of mouth, definition of external edges of mouth and video speech encoding method

It was assumed that the resulting symbols should be in the range from 1 to 50, so the minimum value of the code for each user must be specified accordingly and on this basis the code values need to be reduced to the objective range.

4 The Method of Supporting the Division of Phonemes

The division of the speech signal into phonemes is very important for systems that perform continuous speech recognition. The most often used method currently uses constant-time segmentation. This methods benefit from simplicity of implementation and the ease of comparing blocks of the same length. Clearly, however, the boundaries of speech elements such as phonemes do not lie on fixed position boundaries; phonemes naturally vary in length both because of their structure and due to speaker variations. Constant segmentation therefore risks losing information about the phonemes. Different sounds may be merged into single blocks and individual phonemes lost completely.

Spectral analysis of the speech signal is the most appropriate method for extracting information from speech signals. The analysis of the power in different frequency bands offers potential for distinguishing the start and end of phonemes. Many phonemes exhibit rapid changes in particular sub bands which can be used to detect their start and endpoints. However, for many boundaries, there is no discernible drop in overall power, and at some frequencies, the power is broadly constant over the lifetime of the phoneme.

We propose a method that combines analysis of the frequency signal, and an analysis of the key changes from the video frames of speech. In our approach, we analyze the significant changes in the video frames and synchronize those changes with an acoustic signal. The idea of the described method is shown in Fig. 2

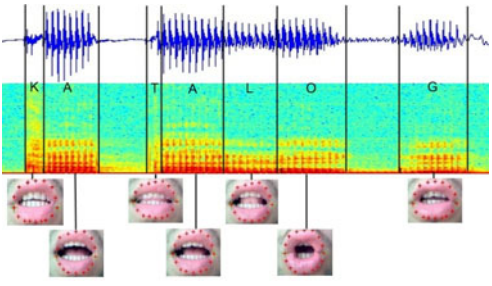

Fig. 2. The method of divide the audio signal into phonemes with use video frame

5 System of Audio Visual Speech Recognition

The accuracy of automatic lip tracking method was tested in audio-visual speech recognition system, that incorporates the hidden Markov models as a probabilistic data classifier. Audio-visual speech recognition is based on the extraction of recording features of audio and features of video. In such system video and audio channels are analyzed separately, then a proper fusion of designated features is made.

The input to HMM has to be presented in the form of vectors of observation. Such an observation vectors can be obtained by making a vector quantization. In audio speech analysis Mel Frequency Cepstral Coefficients (MFCC) were used for the extraction of audio features. To create a codebook Lloyd algorithm was used. In speech recognition systems based on HMM, each frame represented by a vector of observation is coded as a symbol of observation. In our system, all the individual words can be encoded using 37 code symbols, corresponding to the number of phonemes of the Polish language.

The result of audio signal analysis is the extraction of required characteristics of the signal, whereas the result of video analysis being the process of encoding each frame containing the shape of the lips with the use of appropriate symbol observation.

Vectors of observations of audio and video signals have a similar length. The audio signal is sampled at a frequency of 8000 samples per second. After the encoding process one second of audio contains about 50 symbols. Consequently the video signal is sampled at a frequency of 50 frames per second, to synchronize audio and video streams.

In the method of fusion, the audio and visual observation sequences are integrated using a coupled hidden Markov model (CHMM) [11,12]. The feature fusion system using a multi-stream HMM assumes that the audio and video sequences are state synchronous, but allows the audio and video components to have different contributions to the overall observation likelihood. The audio visual product HMM can be seen as an extension of the multi-stream HMM

that allows for audio-visual state asynchrony. The CHMM can model the audio-visual state asynchrony and preserve at the same time the natural audio visual dependencies over time. In addition, with the coupled HMM, the audio and video observation likelihoods are computed independently, significantly reducing the parameter space and complexity of the model compared to the models that require the concatenation of the audio and visual observations [12].

6 Experimental Results

To perform research, we applied a set of seventy-command, recorded for 40 different users. In order to show the correctness of functioning method of automatic tracking of the lips, and the level of error recognition system of audio-video speech, experiments were performed for different levels of audio noise (at SNR of 20, 15, 10, 5, and 0 dB).

Many scientists in the world deal with the analysis of audio-visual speech. In their studies, they examine the various factors of processing audio-visual speech. Therefore, to compare the obtained results with those of other researchers, we chose only those leading the work that analyzed in a similar way audio-visual recognition of speech. In order to compare the developed method with the popular methods of audio-visual recognition of speech, developed by leading researchers in this field, we adopted similar conditions for noisy audio signal. Effectiveness compared with those of [2,13,14,15,16], in which the authors have adopted similar solutions by encoding both signals, and using CHMM for learning and testing. Assumptions may differ in terms of quantity of the analyzed words, different amounts of CHMM states and various means of fusion of audio and video signals. But the sense of studies was similar, so it was concluded that the comparison will be reliable. The results of comparing the level of recognition errors of audio-visual speech was showed in Tab. 1.

Table 1. The results of comparing the level of recognition errors of audio-visual speech

Method	Recognition Accuracy [%]				
	SNR 20dB	SNR 15dB	SNR 10dB	SNR 5dB	SNR 0dB
AV-Concat [2]	88,37	80,66	73,95	66,36	53,73
AV-HiLDA [2]	88,44	81,92	75,91	66,77	56,49
AV-Enhanced [2]	87,28	79,84	70,28	56,88	41,04
AV-MS-Joint [2]	88,63	82,52	77,08	69,97	59,11
AV-LSNR [17]	93,13	88,26	83,05	78,64	71,27
AV-CHMM [18]	98,56	90,08	85,09	75,23	70,51
Audiovisual [19]	94,72	88,16	84,72	74,12	68,96
AV Combined [20]	97,20	93,40	79,50	58,40	50,80
A_Mowa_PL	**96,02**	**77,35**	**52,30**	**37,84**	**29,73**
AV_Mowa_PL	**96,12**	**89,76**	**83,33**	**77,11**	**71,32**

7 Conclusion and Future Work

Conducted tests indicate that the method of automatic lip tracking is working properly and performs well in real life. Test results show that the accuracy of speech recognition is largely affected by the fact of whether the environment is disturbed or not.

In comparison of recognition accuracy our method obtains similar or better results to other existing audio-visual speech recognition methods, published in scientific literature. Fig. 3. shows results of comparing the level of recognition errors of audio-visual speech for different methods.

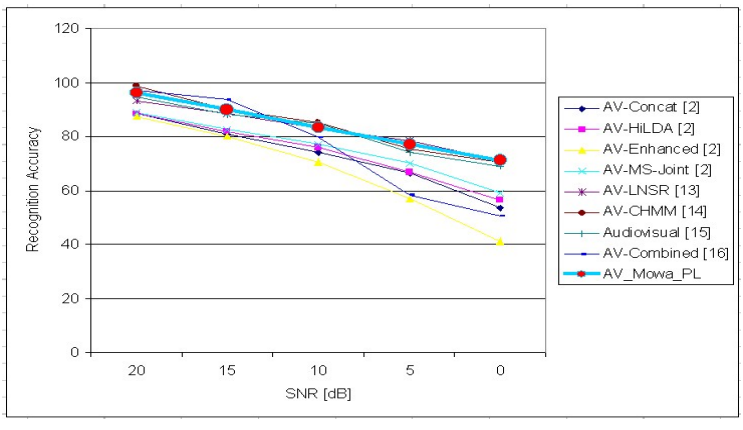

Fig. 3. Results of comparing the level of recognition errors of audio-visual speech for different methods

The results show also that this method should be developed. There are plans to expand the method of automatic detection of the position of the tongue, for each of the spoken video phonemes. In future work we plan to build a system for Polish speech recognition, based on analysis of individual phonemes. Such an approach would allow for continuous speech recognition. The method of audio-visual recognition of Polish speech was used in the system to control the camera movement using voice commands. To increase the efficiency of the method to make the system work properly in real time, can be used to support at the hardware level.

An advantage of the proposed method is the satisfactory effectiveness created by the lip-tracking procedures, and the simplicity and functionality by the proposed methods, which fuse together the audio and visual signals. A decisively lower level of mistakes was obtained in audio-visual speech recognition, and speaker identification, in comparison to only audio speech, particularly in facilities, where the audio signal is strongly disrupted.

References

1. Shin, J., Lee, J., Kim, D.: Real-time lip reading system for isolated Korean word recognition. Pattern Recognition 44, 559–571 (2011)
2. Neti, C., Potamianos, G., Luttin, J., Mattews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., Zhou, J.: Audio Visual Speech-Recognition. 2000 Final Report (2000)
3. Zhi, Q., Kaynak, M.N.N., Sengupta, K., Cheok, A.D., Ko, C.C.: A study of the modeling aspects in bimodal speech recognition. In: Proc. 2001 IEEE International Conference on Multimedia and Expo, ICME 2001 (2001)
4. Jian, Z., Kaynak, M.N.N., Cheok, A.D., Chung, K.C.: Real-time Lip-tracking For Virtual Lip Implementation in Virtual Environments and Computer Games. In: Proc. 2001 International Fuzzy Systems Conference (2001)
5. Petajan, E.: Automatic lipreading to enhance speech recognition. In: Proceedings of Global Telecommunications Conference, Atlanta, GA, pp. 265–272 (1984)
6. Bailly, G., Vatikiotis-Basteson, E., Pierrier, P.: Issues in Visual Speech Processing. MIT Press (2004)
7. Park, S., Lee, J., Kim, W.: Face Recognition Using Haar-like feature/LDA. In: Workshop on Image Processing and Image Understanding, IPIU 2004 (January 2004)
8. Hong, K., Min, J.-H., Lee, W., Kim, J.: Real Time Face Detection and Recognition System Using Haar-Like Feature/HMM in Ubiquitous Network Environments. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3480, pp. 1154–1161. Springer, Heidelberg (2005)
9. Kukharev, G., Kuzminski, A.: Biometric Technology, Part. 1: Methods for Face Recognition. Szczecin University of Technology, Faculty of Computer Science (2003) (in Polish)
10. Chorań, M.: Human Lips as Emerging Biometrics Modality. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2008. LNCS, vol. 5112, pp. 993–1002. Springer, Heidelberg (2008)
11. Kaynak, M.N.N., Zhi, Q., Cheok, A.D., Sengupta, K., Chung, K.C.: Audio - Visual Modeling for Bimodal Speech Recognition. In: Proc. 2001 International Fuzzy Systems Conference (2001)
12. Liu, X., Zhao, Y., Pi, X., Liang, L., Nefian, A.V.: Audio-visual continuous speechr ecognition using a coupled hidden Markov model. In: ICSLP 2002, pp. 213–216 (2002)
13. Hasegawa-Johnson, M., Livescu, K., Lal, P., Saenko, K.: Audiovisual speech recognition with articulator positions as hidden variables. In: Proc. International Congress of Phonetic Sciences (ICPhS) (2007)
14. Shao, X., Barker, J.: Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment. Speech Communication 50, 337–353 (2008)
15. Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K.: A coupled HMM for audio-visual speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP (2002)
16. Nefian, A.V., Liang, L., Pi, X., Liu, X., Mao, C.: An coupled hidden Markov model for audio-visual speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (2002)