An Effective Lip Tracking Algorithm for Acoustic-to-Articulatory Inversion

Jingying Chen, Marie-Odile Berger, Yves Laprie

An effective lip tracking algorithm for acoustic-to-articulatory inversion

Jingying Chen Mario-Odile Berger Yvs Laprie

Lorraine Laboratory for Research into Information Technology and its Applications (LORIA), France

ABSTRACT

Although automatic speech recognition systems can now perform well under certain conditions, they still don't provide good results in real life conditions, especially in noisy environments. Several authors have suggested that using articulatory features rather than acoustic features as a basis for speech parameterization would help yield better recognition results. The articulatory features can be recovered from the speech signal by acoustic-to-articulatory inversion. Given the acoustic signal, the recovery of the articulatory state is considered difficult. The reason is the "one-to-many" nature of the acoustic-to-articulatory inversion problem: a given articulatory state has always only one acoustic realization but an acoustic signal can be the outcome of more than one articulatory states. Since visual information is complementary to acoustic information in the inversion, lip tracking is proposed in this paper to provide visual information of lip movement for the acoustic-to-articulatory inversion. Encouraging results have proven the effectiveness of this method which provides useful information (i.e. mouth width and height) for inversion.

1. INTRODUCTION

Although automatic speech recognition systems can now perform well under certain conditions, they still don't provide good results in real life conditions, especially in noisy environments. Several authors [1] have suggested that using articulatory features rather than acoustic features as a basis for speech parameterization would help yield better recognition results. The articulatory features can be recovered from the speech signal by acoustic-to-articulatory inversion. Given the acoustic signal, the recovery of the articulatory state is considered difficult. The reason is the "one-to-many" nature of the acoustic-to-articulatory inversion problem: a given articulatory state has always only one acoustic realization but an acoustic signal can be the outcome of more than one articulatory states. Since visual information is complementary to acoustic information in the inversion, lip tracking is proposed to provide visual information of lip movement for the inversion in this work.

Works on lip tracking range from purely intensity-based approaches [2] to sophisticated model-based approaches, e.g. active contour models (or snakes [3]) and active shape models [4]. The purely intensity-based approach is simple and fast, however, it is sensitive to the lighting variations. As for the active contour method, it often converges to the wrong result when lip edges are not obvious or when lip color is very close to the face color. Also, the method is quite computationally expensive as they require many iterations to fit the lip contour properly. The active shape model method needs a large set of training data to learn patterns of typical lip deformation [5]. In order to provide useful visual information (such as the width and height of mouth) for the acoustic-to-articulatory inversion, an algorithm using a combination of color and structure information of the mouth area to track the feature points on the inner lip contour is proposed here (see Figure 1). The lip feature points that we are interested in are the two lip corners and the mid-points of upper and lower lips. Since the purpose of the lip tracking algorithm is to help the acoustic-to-articulatory inversion, we are interested in the vocal tract shape so that the inner lip contour is used here.

Figure 1: Lip feature points on the inner lip contour

2. LIP TRACKING ALGORITHM

The lip tracking algorithm uses the combination of color and structure information of the mouth area. Two modules are involved in this approach, i.e. lip localization and lip feature points tracking.

2.1. Lip localization

Before lip feature points tracking begins, the approximate location of the speaker's lips is estimated using color information. First, the image captured from camera is transformed from RGB color space to HSI color space which separates hue (H) and saturation (S) from intensity (I). Then, the hue value is used to calculate the candidate lip pixel because hue is relatively insensitive to the lighting variations. Finally, connected components consisting of pixels with hue value that lies within the range of value typical of the lips are formed, which is identified as the region of interest (ROI) in this study (see Figure 2).

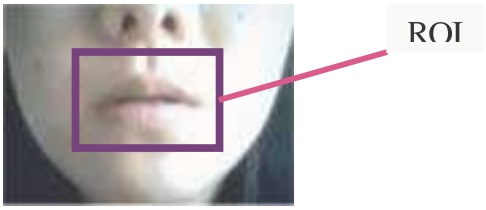

Figure 2: Lip localization.

2.2. Lip feature points tracking

The lip feature points tracking module includes two steps: (1) find the lip corners and (2) find the mid-point of the upper and lower lips.

Step 1: First, a horizontal integral projection [6] is applied on the intensity image in the ROI to find the vertical position of the shadow line between the lips. The shadow line is the darkest horizontally extended structure in the ROI, its vertical position can be found where the horizontal integral projection value is the global minimum (see figure3). The position can be considered as the approximate vertical position of the lip corners.

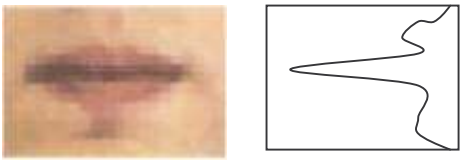

Figure 3: Finding the vertical position of the shadow line between lips using the horizontal integral projection in the ROI.

Second, a vertical integral projection is performed on the horizontal edge map in the ROI. The horizontal edge map can be obtained by applying the Sobel horizontal edge detector. The horizontal positions of the lip corners are estimated by examining the vertical integral projection values, the locations where the values exceed or fall below a certain predefined threshold are considered as the estimated horizontal positions of the lip corners (see figure 4). Then, the vertical position of the lip corners is adjusted by finding the darkest pixel along their horizontal position.

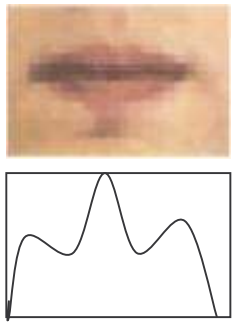

Figure 4: Finding the horizontal position of the lip corners using the vertical integral projection in the ROI.

Finally, the positions of the lip corners are refined. We search for the darkest pixels with maximum contrast around the positions (found above) of the left and right corners in the two small search windows (see figure 5).

Figure 5: Refining the lip corners in the two search windows.

Step 2: The horizontal position of the mid-points of the upper and lower lips is computed as the middle between the left and right lip corners. For the vertical position of the mid-points, two states are considered: closed mouth and open mouth. If the mouth is closed, the mid-points should lie in the shadow line. If the mouth is open, the mid-points should lie either between the teeth and lip flesh or between the oral cavity and lip flesh. Teeth can be separated easily from other parts of the mouth because they have low saturation and high intensity, and oral cavity has low intensity. Hence, the vertical position of the mid-points

can be found using the structure information of the mouth and their color characteristics.

3. EXPERIMENTS

Experiments have been done on different mouth states and different people (from Stirling face database[1]) using C under windows 2000. Some results are given below (see Figure 6 and 7). From these results, one can see that the lip corners and mid-points of upper and lower lips have been detected accurately. According to the positions of these feature points, the mouth width and height can be obtained.

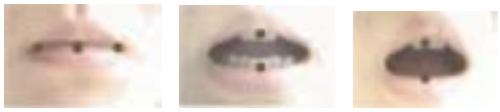

Figure 6: The lip feature points of the different mouth states.

Figure 7: The lip feature points of the different people.

4. CONCLUSION AND FUTURE WORK

A lip feature points tracking algorithm using the color and structure information of the mouth area is presented in this study. Encouraging results have proven the effectiveness of this method. It provides useful information (i.e. mouth width and height) for inversion. Currently, this study is based on 2D images, we are investigating to apply it on stereovision system to provide high accurate 3D mouth position (such as the protrusion of the upper and lower lips).

5. REFERENCE

[1] R.C. Rose, J. Schroeter and M.M. Sondhi, "An investigation of the potential role of speech production models in automatic speech recognition", in Proceedings of the International Conference on Spoken Language Processing, Japan, pp. 575-578, 1994.

[2] J. Yang, R. Stiefelhagen, U. Meier and A. Waibel, "Real time face and facial feature tracking and applications", in Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98, pp 207-212, 1998.

[3] M. Kass, A. Witkin and D. Terzopoulus, "Snake: Active contour models", International Journal of Computer Vision, 1(4), pp 1435-1444, 1992.

[4] J. Leuttin, N.A. Tracker and S.W. Beer, "Active shape models for visual speech feature extraction", Electronic System Group Report No. 95/94, University of Sheffield, UK, 1995.

[5] Y. Tian, T. Kanade and J. Cohn, "Robust lip tracking by combining shape, color and motion", in Proceedings of the 4th Asian Conference on Computer Vision, January, 2000.

[6] T. Kanade, Picture processing by computer complex and recognition of human faces. Technical report, Kyoto University, 1973.

[1] http://pics.psych.stir.ac.uk/cgi-bin/PICS/New/pics.cgi