Extracting Lip Parameters in Speech Synthesis System Driven by Visual-speech

Gang Li, Mengjun Wang, and Ling Lin,
School of Precision Instrument and Opto-Electronics Engineering,
Tianjin University, Tianjin 300072, China
ligang59@tju.edu.cn

Abstract

A speech synthesis system is designed for automatically recognizing visual speech generated by the speech impaired and present a new communication approach for the speech-impaired people. In order to acquire more parameters of lip contours, a new model is made up, which can extract the degree of pouting from it. At the same time, the differential coefficients of some parameters are calculated to describe dynamic characteristic of the lip contours. Movement detection and morphological processing are used to extract mouth area and parameters of lip contours from the image sequence.

1. Introduction

Lip-reading has been an active research area in human-computer interactions in the last few decades. Many researches on it have shown that the ratio of automatic speech recognition could be improved obviously in a noisy environment. So the contents of the speakers are captured or partly captured by tracing the information of lip-movement [1]-[4].

Because of some diseases, some persons' larynxes or vocal cords were resected and they lost the capacity of speech communication. However, these persons have or can be trained to get correct lip movements when they speak. Therefore, it is possible that visual speech information can be automatically recognized and the results can be used to synthesize speech. Inspired by the forenamed achievements, we propose a novel speech synthesis system driven by visual speech. In this system, we focus on automatically extracting as much as possible of visual speech information generated by the speech-impaired people. Then the HMM based recognizer is used to train and recognize the sequence of the visual speech information. The speech synthesis module synthesizes acoustic speech based on the recognition results. This method can improve accessibility and communication for the speech-impaired people.

Lip movements contain most visual speech information, thus visual analysis mainly focuses on exacting the mouth region and the parameters of the lip contours. In the traditional way, the lip contours information have been extracting from the frontal image of the face. But the lip moving is not only in a two-dimensional way, but also in a three-dimensional way [5]. To acquire more information, we adopt a new model to extracting lip-moving information from both the frontal and the profile face at the same time [6]. In this model, both dynamic information of the lip's width, height and the pouting motion are taken into account. Movement detection and morphological processing are used to extract mouth area and parameters of lip contours from the image sequences.

The paper is structured as follows. Section 2 addresses an unsymmetrical lip contours model we use to extract the visual features. Section 3 describes the processing of extracting mouth region and lip contours. Section 4 reports the experimental results on our Chinese visual-speech database for speech-impaired people. Finally, Section 5 summarizes the paper.

2. Unsymmetrical lip contour model

Firstly, we analyzed the features of man's lip when they speaking. The lip contour is eudipleural. When pronouncing a Chinese word, the left-half section and the right-half section of the upper lip or nether lip is moving towards the same direction, up or down. The moving directions cannot be opposite, that is to say, while lift-half of the upper lip is moving upside, the right-half must move towards upside. So, from lift-half of the lip contour can capture the information in the obverse face, and the lip and jaw extruding is a fore-and-aft way with the time changing in the profile face. The model we adopted is shown in Figure 1.

Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06)
0-7695-2616-0/06 $20.00

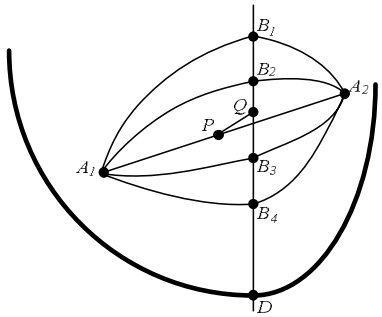

Figure 1. Unsymmetrical lip contours model.

In this model, we make the camera turning an angle away from the vertical direction towards the face. In this model, point A_1 and A_2 are two corners of the mouth, the Euclidian distance of them is the projection of the width of the lip contour, $|A_1A_2| = W$; point B_1, B_2, B_3, B_4 are the midpoint of the outer and inner lip contour, $|B_1B_4| = H$ is the height of the lip contour; In addition, point Q is the midpoint of $|B_1B_4|$, point P is the midpoint of $|A_1A_2|$. From geometrical connection, $|PQ| = E$ is the projection of the distance of the lip extruding, $|PQ|$ reflect the pouting motion.

3. Extracting mouth area

We propose an approach that finds the mouth region and lip contours in the image sequence and then extracts lip parameters. This approach is based upon the follow ideas: when a person is speaking, the human face is quiescent relative to the camera; the lip motion in an image sequence presents high frequency in comparison to other parts of the human face [7]-[9]. Movement detection and morphological processing are used to extract mouth area and parameters of lip contours from the image sequences. The results are not the exact mouth region and lips contours at each frame of the speech, but the features reflect the dynamic information of the lip movement.

3.1 Movement detection

Because mouth region is the main moving object in the image sequences, the mouth region is detected by computing a difference between two consecutive frames in the image sequences [10]. Defining $f_{t-1}(x,y)$ as the original frame at instant $t-1$, $f_t(x,y)$ as the original frame at instant t, the movement is computed as Equation (1).

$$\delta_t(x,y) = |f_t(x,y) - f_{t-1}(x,y)| \quad (1)$$

In the experiment, we find that it is better computing the movement from Equation (2).

$$\delta_t(x,y) = |f_t(x,y) - f_{t-k}(x,y)| \quad (2)$$

k is the number between two frames used. To enhance the movement, an 3×3 averaging mask is used to get the filtered version of $f_{t-k}(x,y)$, and then Equation (2) is written as Equation (3).

$$\bar{\delta}_t(x,y) = | f_t(x,y) - \bar{f}_{t-k}(x,y) | \quad (3)$$

Where, the averaging mask is $\frac{1}{9}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$. In this method, lighting becomes a more feebleness factor.

3.2 Enhancement filtering

To enhance the mouth region and diminish noise, the output frame of the first step is worked by a gray level morphological opening. The opening operation is written as Equation (4).

$$\bar{\delta}_t(x,y) \circ B = (\bar{\delta}_t(x,y) \Theta B) \oplus B \quad (4)$$

Where, B is the structuring element $[1 \quad 1 \quad 1 \quad 1]$.

3.3 Image segmentation

To get the mouth area from the image, we used the maximum entropy thresholding algorithm. In this approach, the 2-D histogram entropies are obtained from the 2-D histogram that is determined by using the gray value of the pixel and the local average gray value of the pixel [11]. It also provides the space relationship between each pixel and its adjacent pixels.

The entropy-based function $\Phi(s,t)$ is defined as the sun of the two entropies. The algorithm then searches for the values of (s,t) that maximizes $\Phi(s',t')$,

$$(s',t') = \arg\arg(\max\Phi(s,t)) \quad (5)$$

In our image sequences, mouth area is the object, others is the background. We get the mouth and lip regions after this step.

3.4 Edge detection

The edge of mouth region got from former steps is the outer lip contours; we extract the edge by morphological processing [12]. A structuring element is used to erode the mouth region, and then the eroded version minus the mouth image, the output is the lip contours. The process is defined as Equation (6),

$$Edge(A) = A - (A \Theta B) \quad (6)$$

Where, A is the image of mouth region from previous steps, B is the structuring element, $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$.

4. Experiments and features extracting

In order to use speech-reading technique into rehabilitation, a mandarin Chinese visual-speech database is designed for the speech impaired. Different from existing visual-speech databases, our database is designed directly for disabled people. It has some specialties as follows: unsymmetrical lip contour model is used to extract the information of pouting; Chinese pronunciation of vowel and consonant is enhanced to improve the dynamic process; considering the developing of lip-reading techniques, this database is fit for expanding. It contains lip's color image sequences when pronouncing 30 usual mandarin Chinese words. These image sequences are collected from ten different students with 320*240, 45*fpt*. Each syllable was read by all the subjects 10 times. Chinese pronunciation of vowel and consonant is enhanced to improve the dynamic process. Then the image sequence for each syllable was segmented into 30 frames long.

These sequences were processed with these steps we put forward. One mouth region and lip contours we got are shown in Figure 2 as an example.

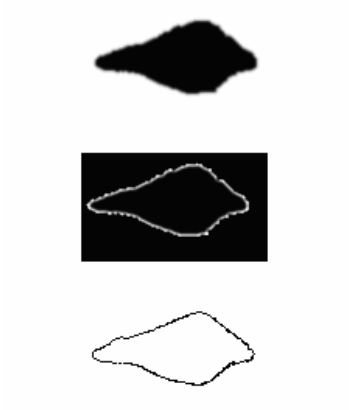

Figure 2. Mouth region and lip contours.

Then, we extract the lip features from the mouth region; including the projection of the width of the outer lip contour W, the height of the outer lip contour H, and the projection of the pouting E. Calculating the maximal distance on the horizontal direction as W, the maximal distance on the vertical direction as H, then find the midpoint (P and Q) of W and H, the Euclidean distance between midpoints is E. The difference of these parameters are calculated as new parameters to describe the dynamic information of the lip, including $\Delta W/\Delta t$, $\Delta H/\Delta t$ and $\Delta E/\Delta t$. Boundary Fourier Transform is used to the lip contour [10] [12]; the Fourier descriptors φ of the lip contour are used as the parameters. These parameters are shown in Figure 3.

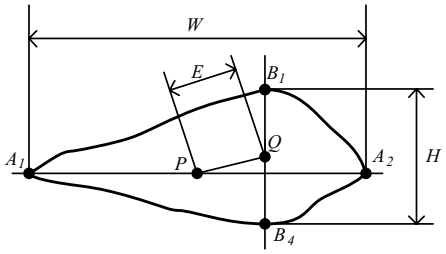

Figure 3. Features of lip contours.

Form the image sequences we extracted and normalized the discrete values of these parameters. The normalized values of E when the same person pronouncing two different Chinese syllables ('Qu' and 'Lai') in different times are shown in Figure 4 as examples, X-coordinates behalf of the frame numbers in image sequences, Y-coordinates behalf of the normalized values of E. Figure 4(a) and Figure 4(b) are the values of E when the same person pronouncing the same syllable (Qu) in different times; Figure 4(c) and Figure 4(d) are the values of E when the same person pronouncing the same syllable (Lai) in different times.

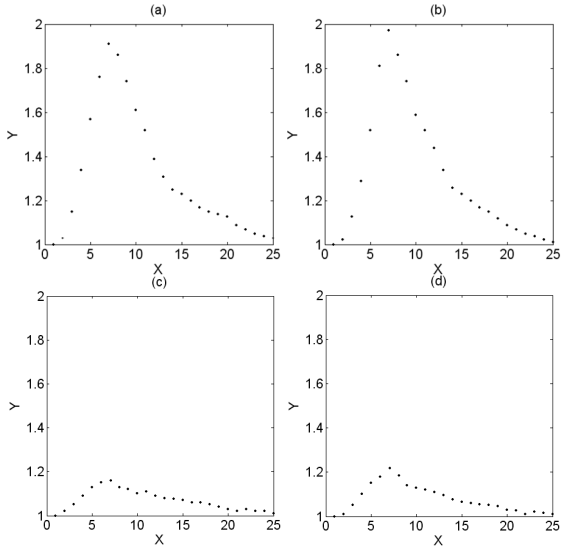

Figure 4. The normalized values of E.

From the results, we can see that the normalized values of E changes in different ways when the same person pronouncing different syllables; and the

normalized values of ***E*** change in a similar way when the same person pronouncing the same syllable.

5. Conclusion

In this paper, a speech synthesis system based on speech-reading technique is investigated for rehabilitation; it presents a new communication approach for the speech-impaired people by using visual information only to synthesize their acoustic speech. We use movement detection and morphological processing extracted mouth area and parameters of lip contours from the image sequence automatically. This approach doesn't need design cost function or pre-define starting control points [1] [2] [7]. From the experiment, we see that the normalized values of ***E*** are changing in different degree in their own image sequences when the same person pronouncing different Chinese words. It can be used as a parameter in our speech synthesis system driven by visual-speech.

6. References

[1]R. Seguier, N. Cladel, "Multiobjectives genetic snakes: application on audio-visual speech recognition", *Proceedings of Fourth EURASIP Conference on Video/Image Processing and Multimedia Communications,* 2003, Vol.2, pp.625–630.

[2]I. Matthews, T.F. Cootes, and J.A. Bangham, "Extraction of visual features for lip-reading", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, Vol.4(5), pp. 198–213.

[3]P. Scanlon, R. Reilly, "Feature analysis for automatic speechreading", *Proceedings of IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp.625–630.

[4]X. Zhang, R.M. Mersereau, M. Clements, and C.C. Broun, "Visual speech feature extraction for improved speech recognition", *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2002, Vol.2, pp.1993–1996.

[5]Z.M. Wang, L.H. Cai, "Study of Chinese Viseme", *Applied Acoustics*, 2002, Vol.21 (3), pp.29–34.

[6]G. Li, M.J. Wang, and L. Lin, "Improving Chinese Lip-reading Recognizing Rate by Unsymmetrical Lip Contour Model", *Optics and Precision Engineering*, 2006, Vol.14(3).(in press).

[7]H.P. Graf, E. Cosatto, and M. Potamianos, "Robust recognition of faces and facial features with a multi-modal system", Proceedings *of IEEE International Conference on Systems, Man, and Cybernetics*, 1997, Vol.3, pp.2034–2039.

[8]W.N. Lie, H.C. Hsieh, "Lips detection by morphological image processing", *Proceedings of IEEE Fourth International Conference on Signal Processing*, 1998, Vol.2, pp. 1084–1087.

[9]L.G. Da Silveira, J. Facon, and D.L. Borges, "Visual speech recognition: a solution from feature extraction to words classification", *Proceedings of XVI Brazilian Symposium on Computer Graphics and Image Processing*, 2003, pp.399 – 405.

[10]Sonka M., Hlavac V., and Boyle R.(Ai H.Z., Wu B. Translated), *Image Processing, Analysis, and Machine Vision (Second edition)*, Posts & Telecom Press, Beijing, 2003.

[11]J.N. Kapur, P.K. Sahoo, and A.K.C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram", *Computer Vision Graphics Image Process*, 1985, Vol.29, pp.273–285.

[12]Gonzalez R.C., Woods R.E.(Ruan Q.Q., Ruan Y.Z. Translated), *Digital Image Processing, (Second edition)*, Publishing House of Electronics Industry, Beijing, 2004.