AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

A System for Automatic Lip Reading

I. Shdaifat and R. Grigat

TU Hamburg Harburg, Vision Systems
D-21071 Hamburg, Germany
{shdaifat,grigat}@tu-harburg.de

D. Langmann

Philips Semiconductors GmbH
D-22529 Hamburg, Germany
detlev.langmann@philips.com

Abstract

In this paper, we present our approach of face and lip detection, lip modeling, and tracking. A new lip model based on Bézier Curves is used to capture the dynamics of the lips efficiently. The model is defined only through few points which are modeled using the Active Shape Model (ASM). Accurate detection of lip details is implemented using multiple independent feature templates. The method detects tracks and model the lips online and robustly. The results of the lip detection are presented based on our collected data for German language. We also describe also some of the current and future work on audio video integration.

1. Introduction

The architecture of an audio-video based speech recognition system is shown by the block diagram (Fig. 1). In this paper, we will elaborate on visual feature extraction, and partially on our current work of the integration of audio and visual features for speech recognition. Our approach comprises the data generation and preparation, face detection, lip modeling, lip detection, and visual feature extraction.

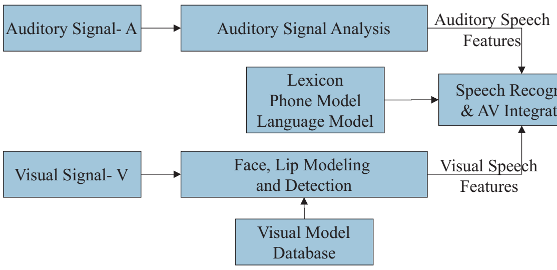


Figure 1: Architecture of an audio-video based speech recognition system

Several methods were proposed for lip modeling and visual speech feature extraction. For example, deformable templates are effective in lip tracking [1]. Kass and Terzopoulos [2] used snakes for lip modeling. Also principal component analysis (PCA) was used [3] and LDA [4]. Basu et al. [5] have introduced a 3D lip model, though near-frontal images comprise useful visual speech data [6]. Active Shape Model (ASM) was presented for lip modeling [7]. Our work is somehow similar to that of [7], we also define the lip boundaries by using parametric curves. In addition, we deploy template matching of several features in order to avoid the noise sensitiveness, which occurs when using intensity data directly [8].

Our new model is able to describe the lips of different speakers using only five Bézier curves that are characterized by nine end and control points. The behavior of the points can statistically be modeled using the *Active Shape Model (ASM)* [9]. We have constructed the German Audio-Visual SAMPA Database. It consists of words containing the phonemes of the German language. We have applied the database to lip modeling and face detection and to sequences from news broadcast in.

This paper is organized as follows: In Sec. 2 we introduce the collected data. In Sec. 3 the detail implementation of face and lip detection is presented. In Sec. 4 the details of the lip modeling is described. The results are presented in Sec. 5. In Sec. 6 the future work of audio video integration is presented.

2. Data Collection

There are some commercially available audio-visual databases [10], and the M2VTS Multimodal Face Database [11] which is a huge collection of images and sound intended for research purposes in the field of multimodal biometric person authentication. The *AVletters* [7] is a collection of isolated letters uttered by 10 speakers of the. Another database is the IBM ViaVoice^{TM} audio-visual database [4]. These databases deal only with English language, and only few efforts have been done for the German Language, e.g. [12].

We have designed our own data set for the German language in which we tried to collect a representative data set of the lip motion covering phonemes and phoneme-to-phoneme transitions. We recorded 57 words form 10 male talkers. Each word represents a phoneme or schwa of the German language, and has been uttered twice. We use the resulting 570 sequences for modeling and recognition tasks. It is necessary to capture near-frontal images of people so that useful visual speech data can be extracted from the images [6]. Therefore, we acquired frontal images of the people. Near frontal images can also be achieved by most of the broadcast news. The broadcast news is a continuous source of sequences which can be easily obtained or recorded. In addition, we have optimal illumination, color, and sound recording conditions. The drawback is that the size of the lips is small for accurate recognition.

3. Face and Lip Detection

The block diagram in Fig. 2 shows our system for face and lip detection. The eye brows are salient features in the face, thus, we use them to detect the face. The skin color is sampled after face detecting, then we restrict the search only in the skin regions for the next video frames. For the first frame in a sequence, we search for the lips using scaled lip templates by the face size obtained from the eyebrow detection. The face detection is decoupled for the next frames and only lip tracking is performed in lip region of interest.

To search for the eyebrows in an image, firstly, we extract all edges, connect them by using short line segments, and then remove edges where the slope angle of the connecting line segment is more than 45^o. The remaining edges are matched with half ellipses. Thereafter, we select the half ellipses pairwise as candidate for the left and right eyebrows, and the symmetry between the left and the right side of the face in addition to the local symmetry of each eye are verified. Finally, we compare the eyebrow pair candidates with left and right eye templates. These eye templates are constructed using our database. The face extraction is illustrated in Fig. 3.

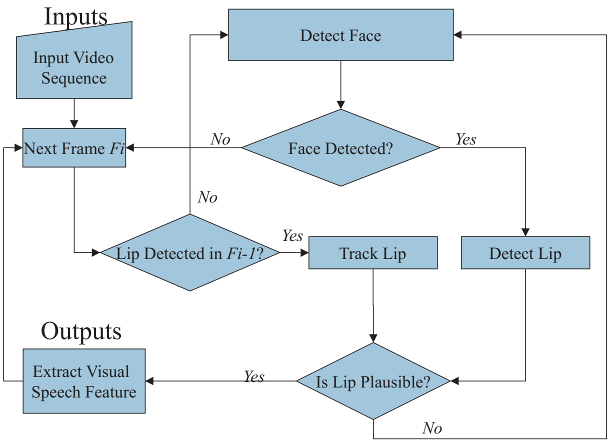
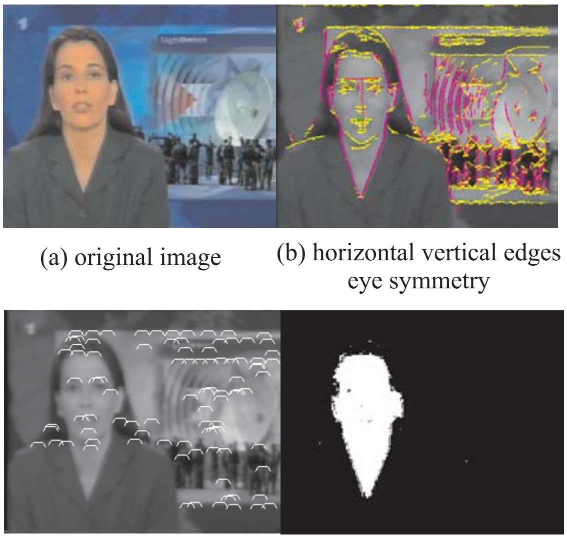


Figure 2: A block diagram of the face and lip detection system


(a) original image

(b) horizontal vertical edges eye symmetry

(c) half ellipses

(d) skin color region

Figure 3: Illustration of face extraction.

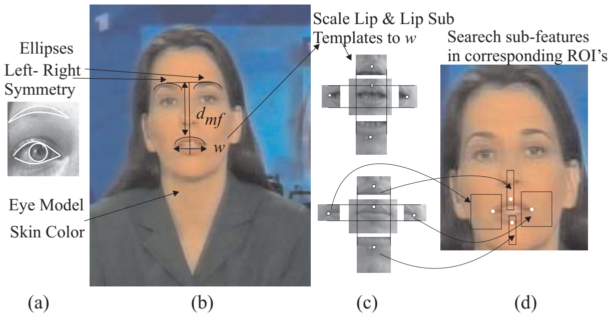

Ellipses
Left- Right
Symmetry
Eye Model
Skin Color
d_{mf}
w
Scale Lip & Lip Sub Templates to w
Seaech sub-features in corresponding ROI's
(a)
(b)
(c)
(d)

Figure 4: Lip and face detection. (a) The features used in face detection. (b) Detection of the face features and finding the lip width. (c) Scaling the lip templates and decomposition into lip sub-templates. (d) Finding the lip features in the corresponding region of interests.

The position of the upper lip is that of the half ellipse under the two eyes (Fig. 3c). We define ζ as the ratio between the lip width w and the distance d_{mf} between the upper lip and the forehead midpoint (Fig. 5). Hence, the width of the lips is roughly

$$w = d_{mf}/\zeta \tag{1}$$

and the lip templates (Fig. 4c) are scaled up to w. Now, we find the principal points of the mouth, which are the two corners of the mouth, the upper and the lower lip. The corners of the mouth are stable features to be detected better than the upper and lower lip. Two regions of interest (ROI) around the two corners of the mouth are calculated using w (Fig. 4d). As we have no idea about the resulting lip shape, we match templates of open and closed lips in the corners ROI. Fig. 4c shows two lip templates of open and closed lips and the decomposition of the lip templates into four lip sub-templates, they are corner of the mouth templates and lower and upper lip templates. Now, we search the corner of the mouth in the calculated region using the resulting corner templates. w is updated with the distance between the two corners. The upper and lower lip templates are scaled with the new w and rotated with the angle between the two corners and the horizon. Two region of interest for the upper and lower lip are calculated and the their templates are matched in these regions.

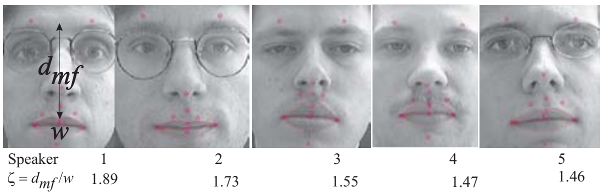


Figure 5: The ratio ζ for different persons.

4. Lip Modeling and Tracking

4.1. Modeling of the lips

We propose a new model for the lip. The model constructed of five Bézier curves (Fig. 6b). Each is defined by two end points $\mathbf{P}_0$, $\mathbf{P}_2$, and one control point $\mathbf{P}_1$ (Fig. 6a), which can be written as

$$\mathbf{P}(t) = \phi_0(t)\mathbf{P}_0 + \phi_1(t)\mathbf{P}_1 + \phi_2(t)\mathbf{P}_2. \tag{2}$$

with $\phi_0(t) = (1-t)^3$, $\phi_1 = 3t(1-t)^2$, and $\phi_2 = (3t^2 - 2t^3)$ and $t \in [0,1]$

The model consists of four end points e_1, e_2, e_3, and e_4, and five control points c_1, c_2, c_3, c_4, and c_5. Where the point e_1 is the position of left corner of the mouth, e_2 is the upper lip, e_3 is the right corner of the mouth, and e_4 is the lower lip, which are the four principal point of the lips. These model compacts 150 feature points defining the lip boundaries where each curve has 30 points (Fig. 6c). The model fits any shape of the German visemes, and is able to capture the dynamics of the lips.

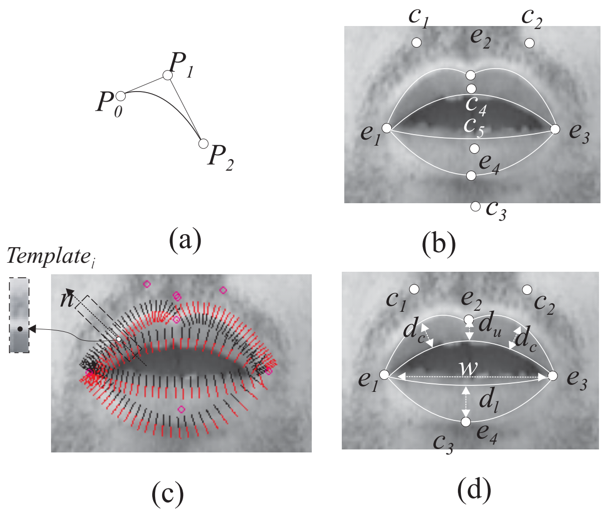


Figure 6: (a) Bézier curve, (b) lip model, (c) textures templates (d) distances d_x used to complete the model from principal points.

The dynamics of the lips can be described by the active shape model (ASM) [7, 13]. We can write a given set of distributed points (x, y) of a certain shape at time i as a vector $\mathbf{v}_i$, where

$$\mathbf{v}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, ..., x_{in-1}, y_{in-1})^T \in \mathbb{R}^n. \quad (3)$$

In our lip model, these are the coordinates of the control points of Bézier curves $e_1, e_2, e_3, e_4, c_1, c_2, c_3, c_4$, and c_5, normalized to d_{mf} (Fig. 5) per talker. Assuming that the positions of the points have a Gaussian probability density function (pdf), the statistics of the distribution of these points can be captured as follows: The mean shape is

$$\overline{\mathbf{v}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}_i. \quad (4)$$

The deviation from the mean is

$$\mathbf{x}_i = \mathbf{v}_i - \overline{\mathbf{v}}. \quad (5)$$

The covariance matrix is given by

$$\mathbf{C} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T. \quad (6)$$

$\mathbf{C}$ can be decomposed using singular value decomposition (SVD)

$$\mathbf{C} = \mathbf{S}\; diag(\sigma_i^2)\; \mathbf{S}^T \quad (7)$$

where $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, ...)$ are the eigenvectors and $(\sigma_1, \sigma_2, ...)$ are the standard deviations within the data along each eigenvector $\mathbf{s}_i$. By Principal Component Analysis (PCA) [13] we can write a shape in the span of the model by

$$\widetilde{\mathbf{x}} = \sum_{i=1}^{M} c_i \sigma_i \mathbf{s}_i = \mathbf{S}\; diag(\sigma_i^2) \mathbf{c} \quad (8)$$

From Equ. 5 the shape of the lip can be approximated by

$$\widetilde{\mathbf{v}} = \mathbf{x} + \overline{\mathbf{v}}. \quad (9)$$

For the purpose of model construction, we start with two extreme cases of the lip shapes as initialization denoted as $\mathbf{v}_1$ and $\mathbf{v}_2$ (Fig. 7). $\mathbf{v}_1$ represents the closed mouth viseme, like the phoneme [*m*] in the word *"Hammer = [ham6]"*(SAMPA notation [14]). $\mathbf{v}_2$ represents the fully opened mouth viseme, like the phoneme [*a:*] in *"Aachen = [a:xn]"*. Other lip shape can be approximated in the span of $\mathbf{v}_1$ and $\mathbf{v}_2$ by

$$\widetilde{\mathbf{v}} = \frac{\mathbf{v}_1 + \mathbf{v}_2}{2} + c\frac{\mathbf{v}_1 - \mathbf{v}_2}{2}. \quad (10)$$

For some speakers, we need an extra template $\mathbf{v}_3$ representing the rounded viseme with rounded appearance like the phoneme [*u:*] in *"Blut = [blu:t]"* in order to obtain a more accurate modeling for that specific speaker. Though, we gained confidence about the significance of $\mathbf{v}_3$ for isolated uttered words, we still need to investigate its importance for the continuous speech. During the training phase, we allowed that single points of the mouth shape may manually be edited, too.

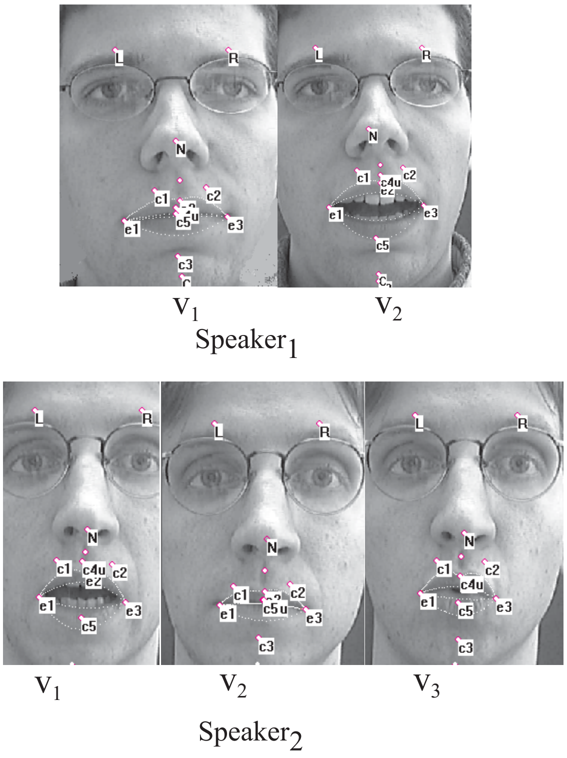


Figure 7: Initial lip templates for two speakers. The feature and Bésier control points have been manually edited

4.2. Completing the Model

After we found the points e_1, e_2, e_3 and e_4 of the lips (see Sec. 3), and constructed the model in the previous section, we can approximate the rest of the shape for a specific speaker as follows: Using the lip width $\overline{e_1 e_3}$ and the lip height $\overline{e_2 e_4}$ we scale the the average $\overline{\mathbf{v}}$ horizontally and vertically to obtain an approximate lip shape $\mathbf{r}$. $\mathbf{r}$ is aligned with the whole model using the method proposed in [13]. The resulting new value

$$\mathbf{x} = \mathbf{S}\; diag(\frac{1}{1+\eta/\sigma_i^2})\; \mathbf{S}^T \mathbf{r}, \tag{11}$$

were the value of η affects the tendency of the shape towards $\overline{\mathbf{v}}$, it is manually chosen where $\eta \in [1, 4]$. and the approximated shape is calculated using Equ. 9. This will correct the point e_4 which is normally difficult to be detected because of the lack of dominant features of the lip in a noisy image.

4.3. Textures of the Shape and Boundary Fitting

For each curve in the model, we selected 30 templates distributed uniformly along each curve (Fig. 6c). Each template is normal to the curve at the point $\mathbf{P}(t_i)$. We fix the endpoints for each Bézier curve, and iterate the control point stepwise until the maximal matching score is reached (Fig. 8). In each step we compute the energy function

$$F = \sum_{i=0}^{30} D_i, \tag{12}$$

where

$$D_i = w_{model}\; \rho_{image,model} + w_{edge}\; \rho_{edge}\; + w_{color}\; \rho_{color}. \tag{13}$$

If the results are consistent and the match score between the model and the lips is above a certain threshold, we only search in a small region of interest (ROI) around each lip feature (Fig. 4d).

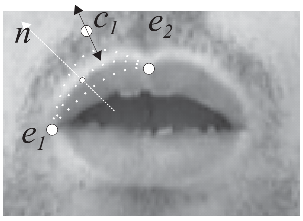


Figure 8: Bézier curve Tracking.

5. Results

We have applied the Bézier model for different persons in our database and on different sequences acquired from the news broadcast. Fig. 9 shows how the model fits the lips of several persons and for different shapes of the mouth. For images of 240 x 320 pixel size, the lip tracking runs about 15 images per second on a 1.5 GHz computer using a webcam, with a speaker-dependent profile. Sequences can be downloaded from [15].

The accuracy of detection and curve fitting was compared to that of hand marked curves for nine sequences. We define the average relative error e between $\widetilde{\mathbf{v}}$ the automatically tracked lip positions and that manually marked lip contours $\mathbf{v}$ relative to the average mouth width $\overline{w}$ in the sequence as

$$e = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{v}_i - \widetilde{\mathbf{v}}_i\| / \overline{w}. \tag{14}$$

Fig. 10 shows e for three speakers of three different image sizes. The error is less than 3% for w $>$ 30 pixel for speaker 2 and 3, whereas for speaker 1, 4 % error has been achieved mainly caused by the moustache.

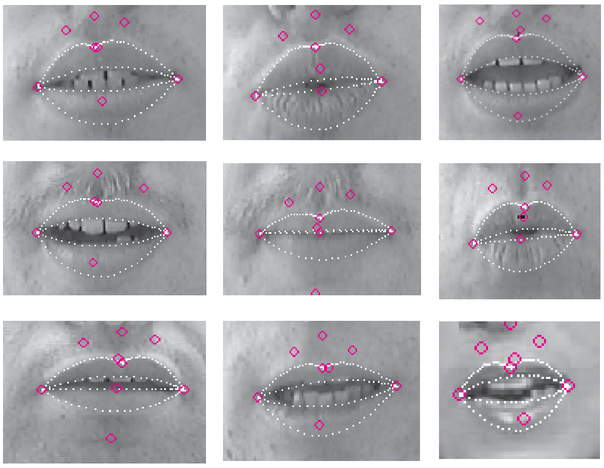


Figure 9: Example of how the model fits the lips.

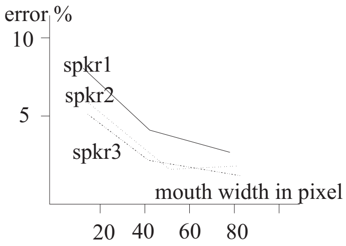


Figure 10: Average relative error between hand-marked and automatically detected shape for different widths of the lips.

Fig. 11 shows the distances $\overline{e_1 e_3} = W$ and $\overline{e_2 c_3} = H$ normalized to the distance d_{mf} vs. time for the three words *"Aachen"*, *"Hammer"*, and *"Blut"*, uttered twice from the German Audio-Visual SAMPA Database. We can notice that, for the phoneme [a:], the mouth hight approaches its peak, whereas, for the phonemes [m] and [b], the valley is reached. For the phoneme [u:] the mouth width decreases remarkably for speakers 2, 5, and 6. A detailed study of these waveforms w.r.t speech is under investigation.

6. Future Work on Audio Video Integration

In this section, we will present the current work on audio video integration. The Hidden Markov Toolkit (HTK) [16] is a widely used free tool for speech recognition and it can be adapted to other tasks. To interface the German language to HTK, we deployed the HADIFIX system and BOMP dictionary [17] in or-

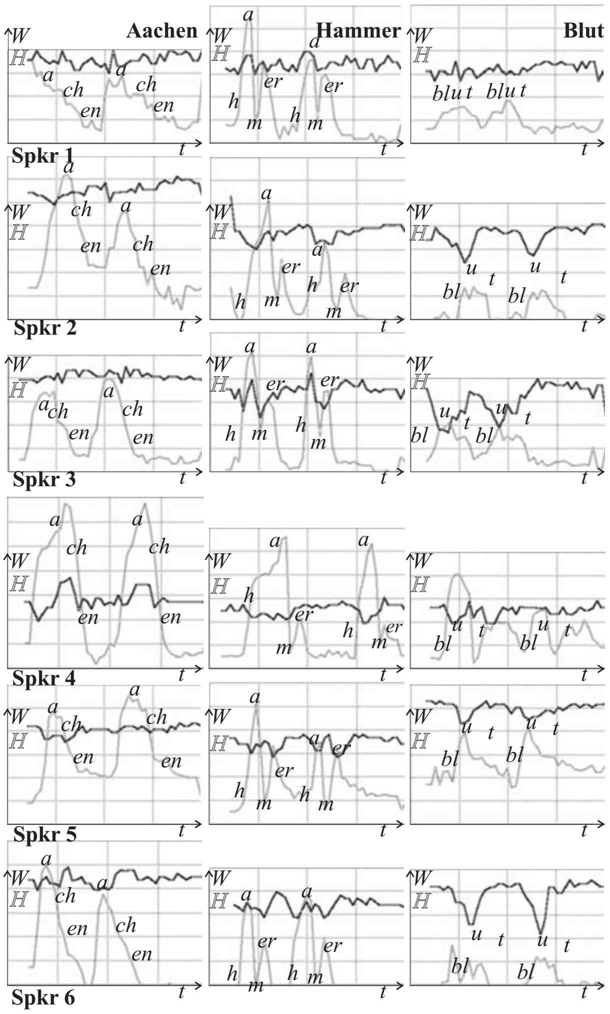


Figure 11: Samples for waveforms of the lip width W (black curves) and height H (gray curves) versus time t for the words Aachen, Hammer, and Blut uttered by 6 speakers. W and H are normalized to the distance d_{mf}

der to convert the text to phonemes. The script and batches can be found in our web site [15].

There are two strategies for audio-video integration in speech recognition the early and the late integration [8]. For simplicity, we currently use only the early integration which is simply input the audio and video feature streams altogether to the recognizer. Existing approaches to visual feature extraction generally fall under two main categories: image-based techniques and explicit feature extraction [8]. We currently extract our lip model parameters and use them with a multi stream HMM. A better way would be using the hybrid approach or a combination of the image pixels and the extracted features of the model, which is currently under investigation. Fig. 12 shows the type of visual speech features which we intend to use. We have done some experiment with those features in [18] where we tried to classify the visemes automatically.

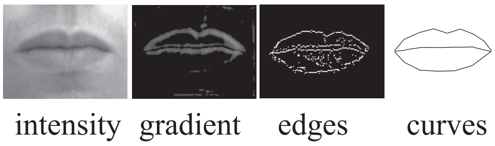


Figure 12: Suggested visual speech features.

7. Acknowledgment

This work is supported by Philips Semiconductors, GmbH, Germany. We would like also to thank the North German Broadcast Corporation - NDR - for supplying us with news broadcast sequences.

8. References

[1] R. Sanchez, J. Matas, and J. Kittler, “Statistical chromaticity models for lip tracking with b-splines,” *In Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication,Lecture Notes in Computer Science, Springer Verlag*, pp. 69–76, 1997.

[2] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” in *International Journal of Computer Vision*, 1988, pp. 321–331.

[3] C. Bregler and Y. Konig, “Eigenlips for robust speech recognition,” *Proc. Int. Conf. Acoust. Speech Signal Processing*, vol. 75, pp. 669–672, 1994.

[4] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, “Large-vocabulary audio-visual speech recognition: A summary of the johns hopkins summer 2000workshop,” in *Proc. Workshop on Multimedia Signal Processing (Special Session on Joint Audio-Visual Processing),*, Cannes, 2001, pp. 619–624,.

[5] S. Basu, N. Oliver, and A. Pentland, “3d modeling of human lip motion,” in *The IEEE International Conference on Computer Vision*, Bombay, India, January 1998, pp. 337–343.

[6] I. Giridharan, N. Chalapathy, P. M. A., Potamianos, and Gerasimos, “Audio-visual data collection system,” U.S. Patent Application Ser. No 20020120643, August 29 2002.

[7] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” *PAMI*, vol. 24, no. 2, pp. 198–213, February 2002.

[8] X. Zhang and C. Broun, “Using lip features for multimodal speaker verification,” in *A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, Greece, 2001.

[9] T. Cootes and C. Taylor, “Active shape models: Smart snakes,” in *BMVC92*, 1992, pp. 267–275.

[10] C. Chibelushi, S. Gandon, J. Mason, F. Deravi, and R. Johnston, “Design issues for a digital audio-visual integrated database,” in *IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication*, London, UK, 28 November 1996, pp. 7/1 – 7/7.

[11] K. Messer, J. Matas, J. Kittler, and K. Jonsson, “Xm2vtsdb: The extended m2vts database,” in *Audio- and Video-based Biometric Person Authentication, AVBPA’99*, Washington, D.C., March 1999, pp. 72–77.

[12] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, "Towards unrestricted lip reading," in *Second International Conference on Multimedia Interfaces*, Hong Kong, 1999.

[13] V. Blanz and T. Vetter, "Reconstructing the complete 3d shape of faces from partial information," *it - Information Technology*, vol. 44, no. 6, pp. 295–302, December 2002.

[14] J. Wells, "Sampa computer readable phonetic alphabet," *URL http://www.phon.ucl.ac.uk/home/sampa/home.htm,Last revised 2003 April 28*, 2003.

[15] "http://www.ti1.tu-harburg.de/˜ shdaifat/."

[16] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge: Entropic Ltd., 1999.

[17] T. Portele, J. Kraemer, and S. Stock, "Symbolverarbeitung im sprachsynthesesystem hadifix," in *Elektronische Sprachsignalverabeitung*, Wolfenbuettel, 1995, pp. 97–104.

[18] I. Shdaifat, R.-R. Grigat, and S. Lütgert, "Viseme recognition using multiple feature matching," *EUROSPEECH, Aalborg, Denmark*, pp. 115–122, September 20001.