Person Verification by Lip-Motion

Maycel Isaac Faraj and Josef Bigun
School of Information Science, Computer and Electrical Engineering (IDE)
Halmstad University, Box 823, SE-301 18 Halmstad
{maycel.faraj, josef.bigun}@ide.hh.se

Abstract

This paper describes a new motion based feature extraction technique for speaker recognition using orientation estimation in 2D manifolds. The motion is estimated by computing the components of the structure tensor from which normal flows are extracted. By projecting the 3D spatiotemporal data to 2-D planes we obtain projection coefficients which we use to evaluate the 3-D orientations of brightness patterns in TV like 2D image sequences. This corresponds to the solutions of simple matrix eigenvalue problems in 2D, affording increased computational efficiency. An implementation based on joint lip movements and speech is presented along with experiments which confirm the theory, exhibiting a recognition rate of 98% on the publicly available XM2VTS database.

1. Introduction

Image sequence based speaker recognition systems have recently attracted research attention [23][17]. The performance of multimodal systems using audio and visual information are known to be superior to those of the acoustic and visual subsystems [3]. Specifically, recognition systems using visual information from lip movements provide supplementary information [4], which can lead to improved speaker recognition performance as demonstrated by [11][5][8][21][22]. The system of [14], used lip contours/shape in each frame in a spatiotemporal image sequence of talking faces and reported performance improvement over acoustic and image based systems. The system requires, however, robust lip contours extraction which is affected by noise, requiring frequent manual intervention. Another disadvantage is the non constant computation time because due to the iterative convergence process of the contours.

This paper describes an algorithm that takes advantage of the low-level spatiotemporal information in an image sequence containing lip-motion. Structure in spatiotemporal images is modeled by moving line patterns in space-time planes, where the normal of the plane encodes the normal velocity of lines. We will present a method for normal velocity estimation based on 3D spatiotemporal space [2] but only by use of 2D signal processing [12]. The lip area is divided into four regions where motion statistics from predefined orientations are extracted for further use in person authentication. This results in increased computational efficiency compared to using the full 3D tensor for normal velocity estimations. Velocities are determined by combining two linear symmetry tensors where each tensor is computed in 2D by cascades of 1D filters. In the next section we discuss normal velocity estimation further. In *section* 3 we present an implementation of the algorithm and the experimental results on data with known ground truth, an expanding circle and a rotating fan with different speed and spatial frequencies. Additionally, in *sections* 4 and 5, we show usefulness of these results by suggesting novel lip movement features for person authentication by reporting results from, to the best of our knowledge, the largest experimental study using both lip-motion and speech features. The full XM2VTS database [16] containing audio and video has been used for performance evaluation, yielding a recognition rate of 98%.

2. Velocity estimation by orientation detection

This section describes the theory for line velocity estimation using orientation detection which we later use to extract visual features for speaker verification. Motion estimation, also known as optical flow, can be determined by eigenvalue analysis of the structure tensor [2]. This method requires multiple time frames since it simultaneously derives the velocities of points and lines. However, for applications that only need line motion features the computations can be excessive. Assuming that only line motion can be observed in local images, the computations can be carried out in 2D subspaces of 3D spatiotemporal space. In lip motion image sequences this assumption is realistic, as our experiments indicate. The original orientation detection in 3D becomes equivalent to a combination of orientation detections in 2D planes. An overview of the spatiotemporal image is pre-

Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)
0-7695-2646-2/06 $20.00 © 2006 IEEE

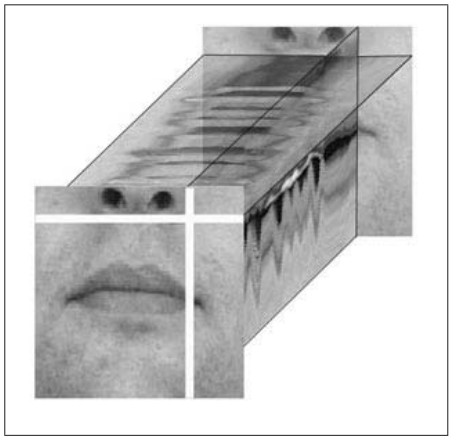

Figure 1. A image sequence from talking faces scene from the XM2VTS database. On the right side of the image-cube, a yt-slice marked by the vertical white line in the xy-image, and on the top of the cube a tx-slice marked by the horizontal line in the xy-image are shown.

sented first and next, the algorithm for velocity estimation is described. Fig. 1 illustrates a space-time image of a moving lip represented as a stack of consecutive 2D images in 3D. At any point we will study the orthogonal cross-sections as regards motion. This is illustrated by the white lines in the figure for a local image point $(x, y, t)^T$. By cutting a cross-section through the cube, we can obtain an xt-slice marked by the horizontal white line. By cutting a cross-section through the cube, we can obtain a xt-slice marked by the horizontal white line. Similarly, we can obtain a vertical cross-section, as marked by the vertical white line in the figure. In the yt cross-section which can be regarded as a 2D space-time manifold, we can see the motion of the lip-movement and mouth orientation in different modes. In the silent mode, we see that the local line directions in the yt-image will be horizontal. By contrast, in the speech mode, the local line-directions will be oblique.

In Fig. 2, we show the ideal situation where an image sequence samples the motion of a line having an arbitrary orientation in the xy-plane. The line motion will appear as inclined lines in both yt- and xt-planes (Fig. 2a). The line motion generates a grey plane in the 3D space-time manifold, xyt, (Fig. 2b). The normal vector of the motion plane is drawn in the figure. We describe now the algorithm which determines the normal velocity of the line-motion from two orientation estimations, in the xt- and yt- manifolds.

The motion of a moving line in a spatiotemporal image generates a plane in 3D but is still a line in the 2D space-time image. The normal velocity in the image plane is determined by the normal of the plane. Assume that the spatiotemporal plane has a normal

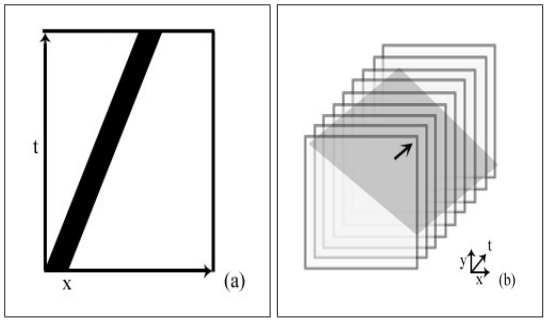

Figure 2. a) A line-motion observed in the 2D space-time manifold. b) A line-motion in the image plane generates the dark plane in the 3D space time image sequence, xyt.

$\nabla \mathbf{f} = (df/dx, df/dy, df/dt)$, also denoted[1] as $\nabla \mathbf{f} = (f_x, f_y, f_t)^T$. The 3D vector $\nabla \mathbf{f}$ is orthogonal to the iso-gray surface of $\mathbf{f}$ at (x, y, t).

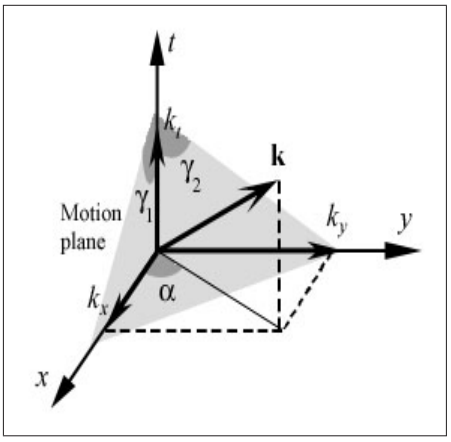

Figure 3. Illustrate a motion plane in 3D space with the original coordinate axes and the corresponding normal axes of the plane in the 3D domain whereas the vector $\mathbf{k}$ is the normal of the plane.

By the shaded plane (Fig. 3), we represent the motion plane together with its normal $\nabla \mathbf{f}$, now replaced by its estimation, $\mathbf{k}$. The angle α represents the direction of the moving line in the image plane. Assuming that the normal of the tilting plane is $\mathbf{k}$, how is the 2D normal flow vector obtained from this 3D vector? In this case we have a linearly symmetric (local) image in 2D, which is an image consisting of iso-gray curves being (parallel) lines.

Assume that the normal of the tilting plane is $\mathbf{k} = (k_x, k_y, k_z)$ where the coordinates x and y represent any arbitrary point in the image plane. By a line movement we

[1]Note that the normal velocity in the xy-plane is 2D and it is invariant to a sign change of the gradient, i.e. the vectors $\nabla \mathbf{f}$ and -$\nabla \mathbf{f}$ represent the same plane, encoding the same velocity vector in the xy plane

have a velocity $v\mathbf{a}$ of a moving line in xyt-space, where $\mathbf{a}$ is the direction of the velocity ($\|\mathbf{a}\| = 1$)

$$\mathbf{a} = (\frac{k_x}{\sqrt{k_x^2 + k_y^2}}, \frac{k_y}{\sqrt{k_x^2 + k_y^2}})^T \quad (1)$$

and v is the absolute speed in the normal direction (in the image plane):

$$v = -\frac{k_t}{\sqrt{k_x^2 + k_y^2}} \quad (2)$$

to the effect that the velocity or the *normal optical flow* will be given by $v\mathbf{a}$
$\mathbf{v} = -v\mathbf{a} = -\frac{k_t}{k_x^2+k_y^2}(k_x, k_y)^T =$

$$-\frac{1}{(\frac{k_x}{k_t})^2 + (\frac{k_y}{k_t})^2}(\frac{k_x}{k_t}, \frac{k_y}{k_t})^T \quad (3)$$

If we know the tilts of the motion plane in the xt- and yt-manifolds, i.e.

$$\tan\gamma_1 = \frac{k_x}{k_t} \qquad \text{and} \qquad \tan\gamma_2 = \frac{k_y}{k_t} \quad (4)$$

we can determine the normal velocity, $\mathbf{v}$. However, the normal of the motion plane, $\mathbf{k}$, is all that is needed to determine the normal velocity.

The tilts $\tan\gamma_1$ and $\tan\gamma_2$ can be estimated optimally in the Total Least Square, TLS, error sense as the local directions of the 2D lines in the xy- and yt-manifolds by using complex convolution, [2] and [1].

$$\tilde{u}_1 = \iint (\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial t})^2 dxdt \quad (5)$$

$$\tilde{u}_2 = \iint (\frac{\partial f}{\partial y} + i\frac{\partial f}{\partial t})^2 dydt \quad (6)$$

It is worth noting that these quantities $\tilde{u}_1$ and $\tilde{u}_2$ are complex valued and that the "tilde" denotes that these are TLS estimations of the true directions. Here f is the 3D image sequence, but the integrations are carried out in 2D planes of the sequence. The obtained complex numbers $\tilde{u}_1$ and $\tilde{u}_2$ correspond to the most significant eigenvectors of the respective 2D structure tensors. They estimate the directions of the lines in the xt- and yt- manifolds, but in the double angle representation, [9]. To be precise, the complex numbers $\tilde{u}_1$ and $\tilde{u}_2$ estimate $2\gamma_1$ and $2\gamma_2$ as follows

$$\tilde{u}_1 = m_1(\cos(2\gamma_1) + i\sin(2\gamma_1)) = m_1\exp(i2\gamma_1) \quad (7)$$

$$\tilde{u}_2 = m_2(\cos(2\gamma_2) + i\sin(2\gamma_2)) = m_2\exp(i2\gamma_2) \quad (8)$$

where m_1 and m_2 are certainty measures. In consequence, the arguments of $\tilde{u}_1$ and $\tilde{u}_1$, must be halved to yield the two tilt angles, γ_1 and γ_2 providing for an approximation of the velocity (3).

$$\tilde{v}_x = \frac{k_x}{k_t} = \tan\gamma_1 = \tan(\frac{1}{2}\arg(\tilde{u}_1)) \quad (9)$$

$$\tilde{v}_y = \frac{k_y}{k_t} = \tan\gamma_2 = \tan(\frac{1}{2}\arg(\tilde{u}_2)) \quad (10)$$

Here, the "tilde" is used again to denote that these quantities are estimations of v_x and v_y.

In our implementation we first used (5)-(6) to compute the two direction angle components needed to obtain the tilts, (9)-(10), which in turn enabled us to estimate the normal image velocities in lip images, via (3). In that, only processing along two planes embedded in 3D spatiotemporal images were needed. In the next section we quantify the accuracy of this motion estimation scheme. We do this by studying the results when the method is applied to synthetic image sequences where the velocities are known.

3. Quantification of motion estimation

In this section we quantify the velocity estimation algorithm by using two synthetic images, a rotated fan and an expanding circle with varying velocities and spatial frequencies.

We implemented the algorithm described above as follows. Let $f(x, y, t)$ be the image sequence.

1. Slice the space-time image f along the vertical and horizontal axis to obtain the tx-image and yt-image sets.

2. Calculate $\tilde{u}_1$ and $\tilde{u}_2$ (using (5) and (6)) at every pixel of the spatiotemporal image.

3. Calculate the velocity $\tilde{v}_x$ and $\tilde{v}_y$ from $\tilde{u}_1$ and $\tilde{u}_2$, according to (9) and (10).

4. Form the complex image sequence pair to represent the velocity.

In the following tests we apply this scheme to quantify the directions and the magnitudes of the normal image velocity estimations. All original images have an intensity dynamic range consisting of the integers in the interval [0,255]. Fig. 4a shows an image containing all possible directions of sine waves with exponentially decreasing frequency in the radial direction of the circles. In the experiments the sine waves were shifted to generate an image sequence (with 64 frames). In Fig. 4b we show the profile along a line (indicated in (a)) where we can observe the spatial frequency in the test image. Fig. 4c illustrates the obtained optical flow estimation for one frame. The length of the arrows represent the magnitude of velocity and the gray-values in

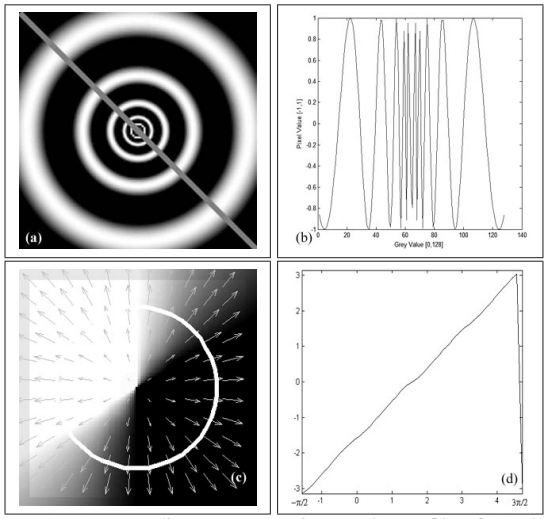

Figure 4. a) Expanding waves test image. b) Profile of (a) along the indicated line. c) The estimated normal optical flow vectors with the orientation estimation in background. d) The graph shows the estimated argument of (c) along the indicated circle.

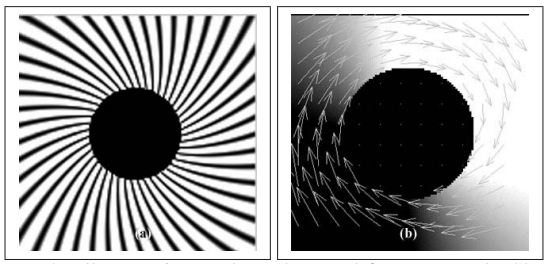

Figure 5. Illustrate in (a) skewed rotated fan pattern. b) Shows the direction and magnitude of *normal image velocity* estimation as arrows and orientation estimation as hue in background.

the background image represent the directions of the estimated velocities. We can see that the gray shift is continuous and monotonous. This velocity direction accuracy is given further precision for the white circle in Fig. 4d, where we observe that the estimated velocity direction follows the true velocity direction very closely since the graph is linear. Additionally, the absolute speeds increase radially in agreement with the ground truth.

The image in Fig. 5a shows a synthetically produced fan. We rotate the image to create artificial motion and obtain 128 frames. The obtained velocity estimation can be seen in Fig. 5b, where the velocity magnitudes and directions are represented by the arrows whereas grey values show dense directions. We can see that the arrow change their directions as the directions in a spiral do. The above results signify a reasonable accuracy of the velocity estimation when local images clearly exhibit line motion. In the following section, we use this velocity estimation in a Gaussian Mixture Model (GMM) framework to perform speaker authentication.

4. Experiments and applications

This section describes the audio, and visual features and studies the role of the visual information in speaker verification. The Hidden Markov Toolkit (HTK) was used to process speech files and preform the GMM analysis [24]. The speech processing for extracting the Mel-Cepstral feature representation is summarized next, for completeness. In *section* 4.1, the image processing for extracting the velocity-statistics based lip image features are presented. Finally, the feature fusion and the Gaussian Mixture Model, used in data-modeling and decision making, are summarized in *section* 4.2.

4.1. Speech analysis and feature extraction

Speech signals in our experiment were recorded at a 16 kHz sampling rate. We extracted a speech frame with the length of 25 ms at every 10 ms. We converted each frame to a 39-dimensional acoustic parameter vector. The vector consisted in 12 cepstral coefficients extracted from the Mel-frequency spectrum [18] of the frame with normalized log energy, 13 delta coefficients, and 13 delta-delta coefficients. The delta and delta-delta coefficients are the first and the second order time derivatives of the extracted cepstral coefficients and are also known as the (speech) velocity and the acceleration respectively.

4.2. Lip-motion features

We are interested in facial changes due to speech production and therefore analyze the mouth region only. Common approaches in face recognition are often based on geometric features or intensity features, either the whole face or part of the face, [6][19]. Much information about the identity of a speaker is contained in the lip-movement and the gray-level changes distribution around the mouth area [7]. Lip reading by optical flow analysis has been shown to be useful [15], because during the speech production the lip deforms and the intensities in the mouth area change in a personal manner. Extracting optical flow features around the mouth area reveals the dynamic information specific to the way the person speaks. The algorithm described in *section* 2 was used to estimate the normal velocities of the lips over each 10 consecutive images of an image sequence for a speaking person. The velocity features were computed in each pixel at the central image frame. Our experiments showed that the most significant motion vectors are around the lip-area which also contains most of the edges.

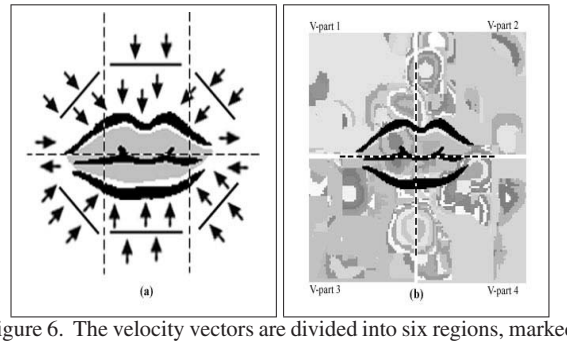

Figure 6. The velocity vectors are divided into six regions, marked by dashed lines where each region is projected into a spatial direction marked by solid line. b) Shows the results of a clustering of the estimated velocity vectors that were divided into four parts by the dashed lines. The gray-values encode the absolute speeds in predefined directions.

4.2.1 Feature clustering

In each mouth-region frame we have numerous points, here 128X128 pixels, with dense 2D velocity vectors. Our goal is to extract statistical features from the normal velocity to reduce the amount of data without degrading identity specific information excessively. First, we reduce the 2D vectors to 1D scalars by only allowing 3 directions $(0°, 45°, -45°)$ as marked with solid lines in 6 regions, (Fig. 6a).The motion vectors within each such region become real scalars that take the signs $+$ or $-$ depending on which direction they move relative their expected directions, the marked solid lines. Then the next step is to quantize these scalar velocities from being allowed arbitrary real scalars to a more limited set of values, here 20. These quantized velocities are obtained from the data by applying an automatic clustering technique, the fuzzy c-means [18], at four regions of the mouth-region (Fig. 6b). The obtained cluster-centers, and their corresponding cluster-populations, were used as a feature vector for each of the 4 regions. In consequence, each of these sub-regions had a 40 dimensional feature vector, consisting of 20 cluster-centers and 20 cluster-populations, summarizing the statistics of lip-motion.

4.2.2 Feature fusion and audio-visual sampling conciliation

The above decribed features are extracted from the raw data (audio and visual speech) and are subsequently combined [20]. In a later step, the acoustic and visual features are merged into a single audio-visual feature vector, as illustrated by Fig. 7. This allows us to develop a joint audio-visual model for person specific information in the data. Furthermore, the recognition methods developed for automatic speech and speaker recognition over three decades can be utilized conveniently. The acoustic and visual sampling rates are, however, different. The speech data as well as visual data are significantly reduced by the feature extraction processing giving an opportunity to synchronize the two data strands at the feature level. As described in the previous paragraph, each visual feature vector corresponds to one fourth of a visual frame (Fig. 7a-b). The factor four is also the rate factor between the audio and the visual data to the effect that we merge each of the four feature vectors of a visual frame with its own audio feature vectors (Fig. 7c). Accordingly, the merged feature vectors come at the rate of the audio feature vectors and have 79 elements, 39 audio and 40 lip-motion. The lip-motion information originating from the same instant are thus distributed in four consecutive samples of the merged data.

4.3. GMM Model

A Gaussian Mixture model (GMM) is Markov Models and therefore we could use the HTK toolkit which was available to us. This kind of model can be understood as a weighted sum of multivariate Gaussian distributions[18].

$$p(\mathbf{x}|\lambda) = \sum p_i b_i(\mathbf{x}) \tag{11}$$

Here $\mathbf{x}$ is a D-dimensional feature vector. A weight p_i represents the mixture weights and the component densities $b_i(\mathbf{x})$ are multivariate gaussian densities. The weights p_i represents the probability that a person identity λ is represented by the feature $\mathbf{x}$ coming from a specific region of the feature space as supported by the guassian b_i. In our system we use subword level (phonemes) using Gaussian Mixture Model (GMM) having 5 states and 3 mixtures in each state.

5. Experiment

This section presents the experimental evaluation of the Gaussian Mixture Model for text-dependent speaker verification. The verification system was evaluated in a task where the speech features and the visual features were combined as in *section* 4. All speaker verification systems built, were based on a subword level (phonomes) using Gaussian Mixture Models. We first present a summary of the XM2VTS database, which is on of the largest publicly available database having both audio and video, and then our experimental results on this database.

5.1. The database

The XM2VTS audio-visual database [16], contains audio and video sequences for 295 speakers (male and female). For each person, several video sequences are taken over four different sessions. The Lausanne protocol [13] or the XM2VTS protocol is a common experimental procedure for speaker verification and identification using differ-

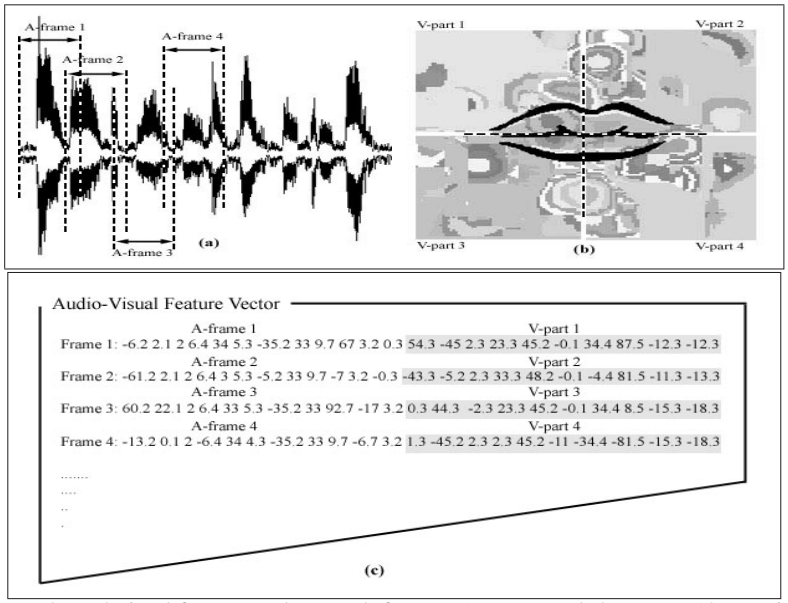

Figure 7. The fusion of speech- and visual-features. The speech features (a) are sampled at every 10 ms with a window width of 25 ms. The clustered visual features (b) are sampled at 25 frames/sec. The fusion, (c), merges every speech feature-vector with one visual feature-vector that only represents the lip-motion of one quarter of an image frame.

ent modalities. The database is divided, using the configuration I in the Lausanne protocol, into different sets, training, evaluation and test sets. Moreover it defines 200 speakers of the 295 as clients, where the first recoding of each sentence of session 1 to 3 is used as training set and the second recording of session 1 to 3 is the evaluation set, and 25 speakers as impostors for evaluation and 70 speakers as impostors for test.

5.2. Experimental results

To investigate the speaker verification performance of the features, the following experiments were conducted on all speakers. The false acceptance rate FA and the false rejection rate FR were calculated as follows

$$FA = \frac{EI}{I} * 100 \quad (12)$$

$$FR = \frac{EC}{C} * 100 \quad (13)$$

Here the number of impostor and impostor acceptance are denoted with I and EI, the number of clients and client rejections with C and EC. Equation (12) and (13) are computed on the evaluation set and the value for which the number of false acceptance and false rejection errors are minimum, i.e. the threshold at which FA rate is approximately equal to FR rate. This threshold of the equal error rate, EER, is used on the test set and is marked as MTER in the figures. The FA is marked by dotted line and the FR is marked by dashed line in the figures. In the first experiment we used a GMM based speech recognition system based only on acoustic features. The threshold function, obtained from the evaluation set (Fig. 8a), is used in the test set to map the score into the confidence interval $[0, 1]$. From the verification results on the test set (Fig. 8b), we obtain the equal error rate of 6%. The verification rate of 295 speakers was thus 94%. In the second test, the system used merged visual and acoustic features on the same basis as the earlier system. The verification rate of the whole database was 98% (Fig. 8d), where the threshold from the evaluation set was utilized (Fig. 8c).

Set / System	Evaluation	Test
Acoustic	96%	94%
Visual	80%	78%
Audio-Visual	99%	98%

Table 1. Verification results on evaluation and test set

To quantify the biometric verification power of the visual features, we carried out the same experiments with these features alone. The FA and FR graphs of the verification results are given in Fig. 9. The verification performance using the threshold of the evaluation set was 78%. In Table. 1 we display the verification performance (where FA is equal to FR in the evaluation set) of the acoustic, visual, and the combined audio-visual systems using the same test data and test protocol. Speaker verification based on audio and visual images from lip-movement give 98% correct classification which is 3-4% better than audio based speaker verification.

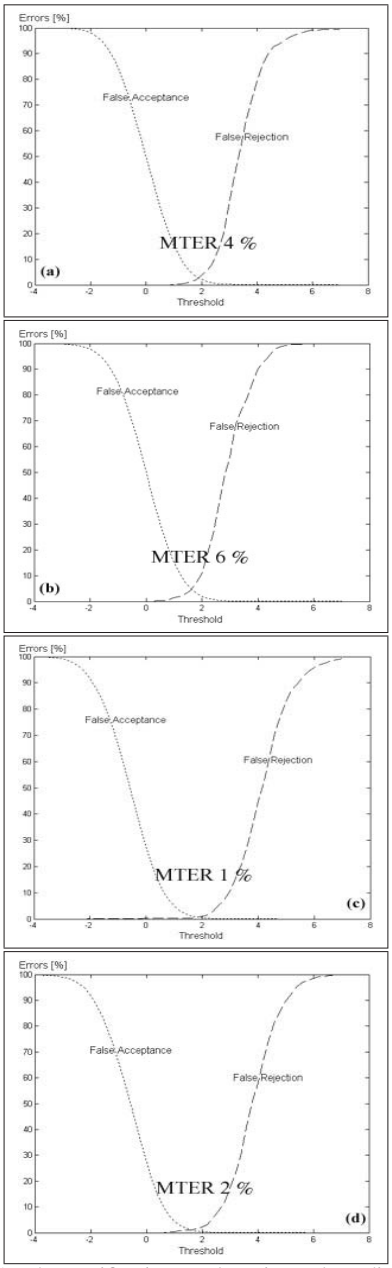

Figure 8. Speaker verification results using only audio signal and audio-visual signal. Acoustic evaluation results are in (a) and verifications (test) results are in (b). The graphs in (c) and (d) show the corresponding evaluation and test results for the combined audio-visual verification system.

This result shows that the audio-visual system achieves better performance than the audio-only system. It is worth noting that the result of the recognition rate in speaker verification is already high (94%) which means that any improvement is difficult. The verification results further show the importance of the visual signal as a complementary information source.

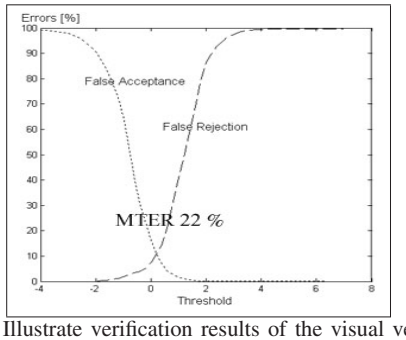

Figure 9. Illustrate verification results of the visual verification system. The threshold is mapped on the evaluation set, where FA rate is equal to FR rate.

5.3. Comparison

We quantified the significance of lip-movements in biometric person authentication as a stand-alone modality as well as in conjunction with the audio modality using a large database. The closest comparable study is the system reported by Jourlin, *et al*. [11]. The experiments in this report, however, were carried out on the M2VTS database, containing 37 speakers. They reported 72% verification rate assuming that the tracking of lip-contours was successful in all frames of all image sequences. We report 78% verification rate on lip-motion only, using a significantly larger dataset, 295 persons. Despite this favorable outcome, the major contribution here consists in that the presented technique does not presuppose a successful tracking since the visual features we suggest require neither segmentation nor lip-tracking.

We can only compare our algorithmic elements with other reports that studied the optical flow for lip-motion, [5][8][21], because these studies have not reported verification results on publicly available audio-visual databases. First, the motion estimation technique used in [5][8][21] measures the motion of points as it avoids regions having the aperture problem. Accordingly, this type of optical flow presupposes the availability of texture in local images, (lack of lines), which is sometimes naturally available, *e.g*. a beard close to mouth, or unshaved male face. By contrast, our motion estimation technique assumes the opposite, namely, it requires the presence of spatial direction, such as lines and edges (not points or isotropic texture), in still image frames. We think that our experiments carried over both female and male data support that these image structures are available in the moving lips themselves in significant amounts in practice. This is all the more significant when considering that the XM2VTS database images consists of entire faces without a particular focus on

the mouth region, causing the lip-areas to have relatively low resolution (128X128). Second, our lip-motion features do not contain any iterative algorithm, as all computations are based on closed form arithmetic function evaluations. The Horn-Schunk algorithm [10], which has been used in the above studies, is iterative in its nature so that the actual computation times per image frame is not constant, making the implementation more difficult on simple computational architectures.

6. Conclusion and discussion

It is experimentally verified that normal velocity estimation in an image is possible by computing a set of directions in 2D projections of the 3D spatiotemporal image-data. The presented results indicate that orientations of projected 2D images yield the normal of an optimal plane which estimates velocities with sufficient accuracy to complement a speaker verification system with lip-motion biometrics. This solution is a computationally efficient alternative to the prevailing velocity estimations, requiring neither segmentation nor lip-tracking by parametric curves.

Visual features are extracted from face image sequences to encode the motion statistics of the lips. Considering the size of the test database, the performance of the system supports the conclusion that lip-motion contains significant dynamic cues for person authentication, yielding approximately 80% verification rates, alone. Due to the low correlation of audio and the motion noise, our motion statistics could improve the performance of even high quality speech (office environment) in speaker verification with additional 3-4 percentage points atop of already high verification rates (approximately 95 %). It is therefore reasonable to conclude that the relative importance of lip-motion to accurate authentication will be higher in noisy or distorted speech.

7. Acknowledgment

We gratefully acknowledge the support of the Swedish Research Council(Vetenskasrdet).

References

[1] J. Bigun. *Vision with direction*. Springer, Heidelberg, 2006. 3

[2] J. Bigun, G. Granlund, and J. Wiklund. Multidimensional orientation estimation with applications to texture analysis of optical flow. *IEEE-Trans Pattern Analysis and Machine Intelligence*, 13(8):775–790, 1991. 1, 3

[3] K. R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995. 1

[4] C. Chibelushi, F. Deravi, and J. Mason. A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 4(1):23–37, 2002. 1

[5] U. Dieckmann, P. Plankensteiner, and T. Wagner. Acoustic-labial speaker verification. pages 301–310, 1997. 1, 7

[6] B. Duc, S. Fischer, and J. Bigun. Face authentication with sparse grid gabor information. *IEEE International Conference Acoustics, Speech, and Signal Processing*, 4(21):3053–3056, 1997. 4

[7] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *The journal of the Pattern Recognition society*, 36(1):259–275, 2003. 4

[8] R. Frischholz and U. Dieckmann. Bioid: a multimodal biometric identification system. *IEEE-Computer Society Press*, 33(2):64–68, 2000. 1, 7

[9] G. H. Granlund. In search of a general picture processing operator. *Computer Graphics and Image Processing*, 8(2):155–173, 1978. 3

[10] B. Horn and B. Schunck. Determining optical flow. *The journal of Artificial Intelligence*, 17(1):185–203, 1981. 8

[11] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. pages 319–326, 1997. 1, 7

[12] K. Kollreider, H. Fronthaler, and J. Bigun. Evaluating liveness by face images and the structure tensor. pages 75–80, 2005. 1

[13] J. Luettin and G. Maitre. Evaluation protocol for the extended m2vts database $xm2vtsdb$, 1998. in: IDIAP Communication 98-054, Technical report R R-21, number = IDIAP. 5

[14] J. Luettin, N. Thacker, and S. Beet. Speaker identification by lipreading. pages 62–65, 1996. 1

[15] K. Mase and A. Pentland. Automatic lip-reading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–76, 1991. 4

[16] K. Messer, J. Matas, J. Kittler, and J. Luettin. Xm2vtsdb: The extended m2vts database. pages 72–77, 1999. 1, 5

[17] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003. 1

[18] D. Reynolds and R. Rose. Robust text-independent speaker identification using gaussian mixture models. *IEEE transactions on Speech and Audio processing* $ICASSP'90$, 3(1):72–83, 1995. 4, 5

[19] M. R. Sanchez, J. Matas, and J. Kittler. Statistical chromaticity-based lip tracking with b-splines. pages 69–76, 1997. 4

[20] J. Stover, D. Hall, and R. Gibson. A fuzzy-logic architecture for autonomous multisensor data fusion. *IEEE Transactions on Industrial Electronics*, 43(3):403–410, 1996. 5

[21] S. Tamura, K. Iwano, and S. Furui. Multi-modal speech recognition using optical flow analysis for lip images. *Journal of VLSI Signal Processing*, 36(2):117–124, 2004. 1, 7

[22] X. Tang and X. Li. Fusion of audio-visual information integrated speech processing. pages 127–143, 2001. 1

[23] X. Tang and X. Li. Video based face recognition using multiple classifiers. pages 345–349, 2004. 1

[24] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The htk book (for htk version 3.0), 2000. http://htk.eng.cam.ac.uk/docs/docs.shtml. 4