ICA-BASED LIP FEATURE REPRESENTATION FOR SPEAKER AUTHENTICATION

S.L.Wang[†] and A. W. C. Liew[#]

[†]School of Info. Security Engg., Shanghai Jiaotong University, Shanghai, China
[#]School of Info. and Comm. Technology, Griffith University, Brisbane, Australia

ABSTRACT

Compared with some “static” biometrics such as human face and fingerprint, person authentication based on lip movement has the advantage of incorporating “dynamic” features which contain rich information indicating the speaker identity. This paper proposes a new lip feature representation and analyzes its discrimination power for person authentication. Since the original lip features are usually of high-dimension, the independent component analysis (ICA) is adopted for dimension-reduction and discriminative feature extraction. Hidden Markov Model (HMM) is then employed as the classifier for its superiority in dealing with time-series data. Experiments are carried out on a database containing 40 speakers in our lab. By analyzing the experimental results, detailed evaluation for various lip feature representation is made and 98.07% accuracy rate in speaker recognition and 2.31% EER in speaker authentication is achieved using our lip feature representation.

Index Terms — lip feature, ICA, speaker authentication

1. INTRODUCTION

Recent research has shown that using the biometric features of a person for authentication can provide much greater security than simply using the conventional password or Personal Identity Number (PIN) accessing method. Biometric features such as fingerprint, iris, face, and hand have been proposed and used for person authentication in many security systems. Visual information about the lip movement has recently been used as a new biometric feature in a multimodal person authentication system.

Various researches have proposed a variety of techniques of using lip feature for speaker authentication/ verification [1-3]. Brown et al. [1] extract identity-relevant information from six geometric lip features and two inner mouth features indicating the visibility of the teeth and the tongue. Polynomial-based approach is adopted for classification. Wark and Sridharan [2] take profiles along normals to the contour points and concatenate them to form the grand profile vector (GPV) for the image. LDA-PCA features extracted from the GPV are employed as the lip features and Gaussian Mixture Model (GMM) is adopted to model the uttering style for a specific speaker. Luettin et al. [3] employ the Active Shape Model (ASM) to model the lip shape and extract the ASM features and the intensity profile vector along normals to the model points. A 3-state HMM is used for classification. As a result, all the lip features used for speaker authentication/verification in the literature can be categorized into two kinds: shape features and intensity variation features around the lip contour (contour features in short). Recently, Matthews et al. [4] have proposed a new kind of feature considering the intensity variation inside the outer lip contour (referred as the lip texture feature), which has shown effectiveness in visual speech recognition. In this paper, the texture feature is also introduced for speaker authentication. Since the above features are usually of high dimension with much redundant information, PCA has been used for dimension-reduction in the literature [3, 4].

Independent Component Analysis (ICA) [5] aims to extract statistically independent components from the original signal. In this paper, it is adopted for dimension reduction and discriminative feature extraction from the three types of lip features mentioned above. Finally, a six-state HMM is employed to model the moving lips of a specific speaker. A database containing 40 speakers in our lab is built and experiments are carried out to evaluate the authentication performance using both the ICA-based lip feature representation compared with that of many other feature representations appeared in the literature.

The paper is organized as follows. In section 2, the previous work of our group is introduced including the lip contour extraction and lip feature selection. Section 3 presents the derivation of all the shape features, contour features and texture features from the extracted lip contour. The ICA-based feature extraction method is also introduced in this section. Authentication results are shown in section 4 using various lip feature representations and section 5 draws the conclusion.

2. PREVIOUS WORK

2.1 Lip region segmentation and lip contour extraction

In order to extract the accurate lip contour in an efficient manner, a two-stage lip contour extraction algorithm is adopted. In the first stage, the FCMS algorithm

The work described in this paper is supported by the Shanghai Natural Science Fund (05ZR14080), the NSFC fund (60702043) and the National Key Technology R&D Program (2006BAK13B09).

978-0-7695-3122-9/08 $25.00 © 2008 IEEE
DOI 10.1109/SITIS.2007.37

[6] is employed to partition all the pixels into two categories: lip pixels and non-lip ones. Considering the presence of image noise and ambiguity, a fuzzy segmentation process is performed and then a membership map is generated which assigns a lip-class probability value to each pixel (shown in Fig.1 (a) and (b)).

With the membership map, our previously proposed point-based lip contour extraction method [7] is adopted to derive the lip outer contour. A 16-point lip model is employed to describe the lip contour and some geometric lip constraints are applied to guarantee the valid lip shape. Then a cost function is formulated to maximize the lip probability and minimize the non-lip probability inside the extracted lip contour. The final lip contour is obtained by a point-driven optimization technique in an efficient manner.

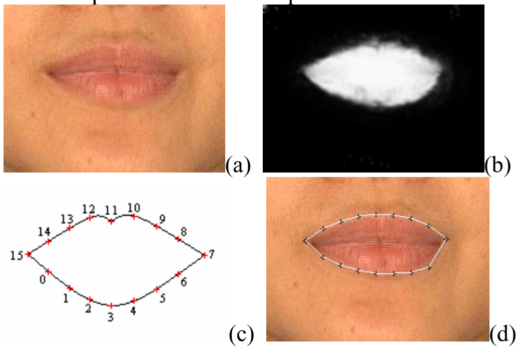

Fig. 1 (a) original lip image; (b) membership map of (a); (c) geometric lip constraint model; (d) lip contour fitting result.

2.2 Lip feature selection

In [8], we have proposed two kinds of lip features for speaker authentication: the ASM features which describe the lip shape information and the intensity profile along the horizontal and vertical line of the lip which describe the intensity variation. Such kinds of lip features can provide discriminative information for speaker authentication [8] and they are employed in this paper for comparison.

3. DISCRIMINATIVE LIP FEATURE EXTRACTION

3.1 Speaker-relevant lip characteristics derivation

With the lip outer contour obtained, three kinds of speaker-identity related lip characteristics: the shape-based, contour-based and texture-based features are extracted.

3.1.1 Shape-based lip information

The width and height of the lip are the most widely used geometric lip features for visual speech recognition and visual speaker verification. They can be obtained by calculating the Euclidean distance between point 7 and point 15 and the Euclidean distance between point 3 and point 11, respectively (shown in Fig.1 (c)). In order to reduce the variance caused by different camera setting of the user, the geometric features are normalized by dividing the width/height value of the first image.

Besides the geometric lip features such as the width and height of the lip, the relationship among the lip contour points can also provide useful information to indicate speaker identity. Due to the variation in translation, scaling and rotation, the coordinates of the contour-points cannot be used as shape-based features directly. To avoid such effects, the shape alignment scheme in [8] is adopted for contour-point feature normalization. As a result, the shape-based lip information is represented by $\{\mathbf{f}_{\mathbf{geo}}, \mathbf{f}_{\mathbf{shape,raw}}\}$, where $\mathbf{f}_{\mathbf{geo}}$ represents the normalized geometric features and $\mathbf{f}_{\mathbf{shape,raw}}$ represents the normalized contour coordinates.

3.1.2 Contour-based lip information

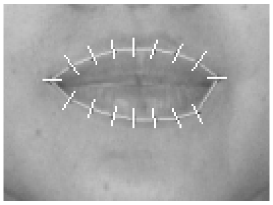

Fig. 2 Normals to the contour points in the intensity map

The contour-based lip information is composed of a set of intensity profiles along the normals to the contour points (shown in Fig. 2). Such information is widely used in many lip contour extraction methods and is shown to be useful for speaker recognition [3]. The length of the profile vector is selected to be large enough (21 pixels in our experiments) to cover both the skin area around the lip and the oral cavity (for open mouth). The final feature, $\mathbf{f}_{\mathbf{contour}}$, is derived by concatenating all the contour point profiles in a preset order (from point 0 to point 15). Note that the magnitude of the original feature vector largely depends on the lighting condition while its variation contains much information about the lip. Hence, the mean-subtraction technique is adopted as a preprocessing for extracting contour-based features.

3.1.3 Texture-based lip information

The texture-based lip information considers the intensity distribution inside the outer lip contour, which may contain the lip, teeth, tongue and oral cavity. To avoid the variations caused by scaling, rotation, various camera settings and lighting condition, two pre-processing methods, shape alignment and intensity normalization are performed.

Shape Alignment: For each lip image being processed, the extracted contour points $\mathbf{x}_{\mathbf{s}}$ and the mean lip shape $\overline{\mathbf{x}_{\mathbf{s}}}$ are aligned and the entire region inside the outer lip contour is then projected onto the mean lip shape. Detailed shape alignment method can be found in [9].

Intensity Normalization: After shape alignment, the intensity distribution inside the lip contour is projected onto the same reference lip shape, i.e., the mean shape $\overline{\mathbf{x}_{\mathbf{s}}}$, and thus the variation caused by different lip shape is avoided.

Then an iterative approach is employed to derive a reference lip texture distribution for intensity normalization, which runs as follows:

1. Project the intensity distribution for all the lip images in the training set onto the reference lip shape and form the shape-normalized intensity distribution vectors $\{\mathbf{I_1}, \mathbf{I_2},\ldots,\mathbf{I_{400}}\}$ (400 training samples in all).
2. Set the initial value of the reference texture distribution $\mathbf{I_{ref}}$ as $\mathbf{I_1}$**.**
3. Normalize each the intensity distribution vector $\mathbf{I_i}$ (i=1,2,…,400) with respect to the reference texture distribution $\mathbf{I_{ref}}$ by [4],
$$\mathbf{I_{nor,i}} = (\mathbf{I_i} - mean\_i \cdot \mathbf{1})/amp \qquad (1)$$
$$mean\_i = \mathbf{I_i} \cdot \mathbf{1}/m\,, \quad amp = \mathbf{I_i} \cdot \mathbf{I_{ref}} \qquad (2)$$
where *mean_i* is the average intensity value of $\mathbf{I_i}$, *m* is the number of elements in the vector $\mathbf{I_i}$.
4. Derive $\mathbf{I_{ref,new}}$ by the mean value of the normalized intensity distribution vector, i.e.,
$$\mathbf{I_{ref,new}} = \frac{1}{400}\sum_{i=1}^{400}\mathbf{I_{nor,i}} \qquad (3)$$
5. Repeat step 3 and 4 until converge, i.e., the Euclidean distance between $\mathbf{I_{ref}}$ and $\mathbf{I_{ref,new}}$ is less than a preset threshold ε.

Fig.3 shows the converged reference intensity map obtained from the training set. For any testing data, the normalized intensity distribution vector can be derived by eqn. (1) and (2).

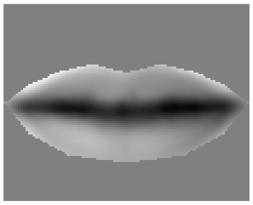

Fig.3 The reference intensity distribution map

3.2 Discriminative feature extraction by ICA

Since the speaker-relevant lip information mentioned above may be of relatively high dimension and sensitive to the accuracy of the contour extraction results, the independent component analysis (ICA) is employed to extract more robust features representing the lip shape. Independent component analysis is a technique to decompose the non-gaussian data into several statistically independent components. Such a representation seems to capture the essential structure of data in many feature extraction applications [5].

To extract low-dimensional, essential feature by ICA, the original data is assumed to be a linear mixture of an unknown set of N statistically independent source shapes, i.e.,

$$\mathbf{f_{raw}} = a_1\mathbf{s_1} + a_2\mathbf{s_2} + \ldots + a_N\mathbf{s_N} = \mathbf{a_{data}}\mathbf{S_{data}} \qquad (4)$$

where N independent sources $\mathbf{s_i}$ (i=1,2,…,N) form the row of source matrix $\mathbf{S_{data}}$.

In order to derive the source matrix $\mathbf{S_{data}}$ from the training data, the fastICA algorithm [5] that maximizes the statistical independence between estimated sources is used. Hence, the shape-based ICA lip feature is represented by,

$$\mathbf{f_{ICA}} = \mathbf{a_{data}} = \mathbf{f_{raw}} \cdot \mathbf{S}^T_{\mathbf{data}} \cdot (\mathbf{S}^T_{\mathbf{data}}\mathbf{S_{data}})^{-1} \qquad (5)$$

4. EXPERIMENTAL RESULTS

Due to there is few public, high-quality lip sequence database available, a database consisting of 40 speakers with 29 males and 11 females in our lab uttering the same phrase three-seven-two-five (3725) in English is built. Each speaker was asked to repeat the phrase for ten times and each utterance contains 90 lip images lasting for 3 seconds. The lip image is with relatively high resolution of 110×90 pixels describing the mouth region. A left to right continuous density Hidden Markov Model (HMM) with diagonal covariance matrix Gaussian model associated with each state is adopted as the classifier. To select proper HMM parameters, the performances for various number of Gaussian mixtures (1, 2, 3 and 4) and number of states (5, 6, 7 and 8) have been investigated. The Baum-Welch algorithm following the Maximum Likelihood (ML) criterion have been used for training the HMM, and the Viterbi algorithm for recognition. Experimental results show that the left to right HMM with six states, two continuous density with diagonal covariance matrix Gaussian mixtures associated with each state delivers the best performance.

To comprehensively analyze the authentication performance, both recognition and authentication tests are employed. For recognition tests, three sets of data are used to train the speaker model for each individual and the remaining seven sets for the same speaker are used for testing. The testing lip sequence is recognized as the speaker HMM with maximum likelihood and the recognition accuracy is adopted to evaluate the performance using various kinds of lip features. For authentication tests, three sets of data for all the speakers are used to train the "world model" for the similarity measure. The similarity score is calculated by taking the log likelihood ratio between the scores of the speaker model and "world model". And in the experiment, all the lip sequences not belonging to the subject are regarded as the imposter data. For each utterance, when its similarity score is greater than a preset threshold, it will pass the verification and vice versa. The proper threshold is selected when the false accept rate (FAR) equals to the false reject rate (FRR) and thus the equal error rate (ERR) is obtained for feature evaluation. It is obvious that the selection of training samples will affect the authentication performance especially when the number of training samples is limited. For better statistical validity, one hundred random tests have been performed for each test and the average performance is taken as the final accuracy/error rate for the specified feature set.

To evaluate the effectiveness of the ICA-based lip feature representation for speaker authentication, several widely-used lip descriptors in the literature are employed for comparison, including: the original shape features {$\mathbf{f_{geo}}$, $\mathbf{f_{shape,raw}}$} (introduced in Section 3.1.1), the PCA-based shape features (similar to the ASM features in [8]), the PCA-based contour features (similar to the intensity profile in [3] and [8]) and the PCA-based texture features (similar to the AAM features in [4]).

It should be noted that the dimension of the shape/contour/texture feature for PCA and ICA is set to (4,4), (40,30) and (300,100) empirically which provides better authentication result compared with other settings. Table 1 and Table 2 demonstrate the recognition accuracy and authentication equal error rate for each feature set and their combinations.

Features	Recognition Accuracy (%)	Authentication EER (%)
$\mathbf{f_{geo}}$	70.23	16.37
$\mathbf{f_{shape,raw}}$	82.27	9.62
$\mathbf{f_{shape,PCA}}$	83.69	9.07
$\mathbf{f_{shape,ICA}}$	83.61	9.14
$\mathbf{f_{contour,PCA}}$	84.88	8.77
$\mathbf{f_{contour,ICA}}$	84.35	8.51
$\mathbf{f_{texture,PCA}}$	93.92	4.03
$\mathbf{f_{texture,ICA}}$	**95.74**	**3.68**

Table. 1 Speaker recognition and authentication accuracy in % by HMM with different kind of single feature.

Features	Recognition Accuracy (%)	Authentication EER (%)
$\mathbf{f_{geo}}+\mathbf{f_{shape,PCA}}$	89.42	5.91
$\mathbf{f_{geo}}+\mathbf{f_{shape,ICA}}$	89.68	6.04
$\mathbf{f_{geo}}+\mathbf{f_{contour,PCA}}$	90.13	5.84
$\mathbf{f_{geo}}+\mathbf{f_{contour,ICA}}$	89.92	6.01
$\mathbf{f_{geo}}+\mathbf{f_{texture,PCA}}$	95.19	3.33
$\mathbf{f_{geo}}+\mathbf{f_{texture,ICA}}$	97.04	2.97
$\mathbf{f_{geo}}+\mathbf{f_{shape,PCA}}+\mathbf{f_{contour,PCA}}$	93.96	4.12
$\mathbf{f_{geo}}+\mathbf{f_{shape,ICA}} + \mathbf{f_{contour,ICA}}$	94.17	3.81
$\mathbf{f_{geo}}+\mathbf{f_{shape,PCA}}+\mathbf{f_{texture,PCA}}$	96.74	2.88
$\mathbf{f_{geo}}+\mathbf{f_{shape,ICA}} + \mathbf{f_{texture,ICA}}$	**97.92**	**2.37**
$\mathbf{f_{geo}}+\mathbf{f_{contour,PCA}}+\mathbf{f_{texture,PCA}}$	95.24	3.28
$\mathbf{f_{geo}}+\mathbf{f_{contour,ICA}} + \mathbf{f_{texture,ICA}}$	96.92	3.02
$\mathbf{f_{geo}}+\mathbf{f_{PCA,all}}$	96.77	2.82
$\mathbf{f_{geo}}+\mathbf{f_{ICA,all}}$	**98.07**	**2.31**

Table. 2. Speaker recognition and authentication accuracy in % by HMM with different combination of lip features.

The following observations may be made from the experimental results:

1. The authentication performance substantially increases using the shape, contour intensity variation and mouth intensity distribution information compared to the geometric lip information, which indicates the intensity variation information contains more discriminative information compared to the shape of the lip outer contour.
2. For the texture-based features, the geometric and shape features can provide additional information indicating the speaker identity, while the contour-based features are redundant.
3. Compared with PCA, ICA-based feature representations obtain similar results for shape and contour features while show better performance for the high-dimension texture-based features.

Considering all the three kinds of lip features, the authentication system can obtain 98.07% accuracy for speaker recognition and 2.31% EER for authentication.

5. CONCLUSIONS

This paper proposes a new lip feature representation for speaker identity authentication, which contains the shape-based, contour-based and texture-based features. The independent component analysis (ICA) to extract speaker discriminative lip features is presented and its authentication performance is analyzed in detail compared with the raw data and PCA. HMM is employed as the classifier and the experimental results demonstrate: i) the texture information contains much identity-related information compared with the shape and contour information; ii) exploiting all the lip features including the geometric features and the ICA-based features can derive high authentication performance, i.e., 98.07% accuracy rate in speaker recognition and 2.31% EER in speaker authentication in our experiment.

6. REFERENCES

[1] C. C. Brown, X. Zhang, R. M. Mersereau and M. Clements, “Automatic speechreading with application to speaker verification”, *Proc.* 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (*ICASSP’02*), vol.1, pp. 685-688, Orlando, USA, May 2002.

[2] T. Wark and S. Sridharan, “A Syntactic Approach to Automatic Lip Feature Extraction for Speaker Identification”, *Proc.* 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (*ICASSP’98*), vol.6, pp. 3693-3696, Seattle, USA, May 1998.

[3] J. Luettin, N. A. Thacker and S. W. Beet, “Learning to recognise talking faces”, *Proc.* 13th International Conference on Pattern Recognition, vol.4, pp. 55-59, Vienna, Austria, Aug. 1996.

[4] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, R. Harvey, “Extraction of visual features for lipreading”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.24, issue 2, pp. 198-213, Feb. 2002.

[5] A. Hyvarinen, E. Oja, “Independent Component Analysis: Algorithms and Applications”, *Neural Networks*, vol. 13, pp. 411-430, 2000.

[6] S.H. Leung, S.L. Wang and W.H. Lau, “Lip image segmentation using fuzzy clustering incorporating an elliptic shape function”, *IEEE Trans. on Image Processing*,

vol.13, issue 1, pp.51-62, Jan. 2004.

[7] S. L. Wang, W. H. Lau and S. H. Leung, “Automatic lip contour extraction from color images”, *Pattern Recognition* , vol.37, no.12, Dec. 2004.

[8] L. L. Mok, W. H. Lau, S. H. Leung, S. L. Wang and H. Yan, "Person Authentication Using ASM Based Lip Shape and Intensity Information", *Proc. IEEE International Conference on Image Processing (ICIP 2004)*, Singapore, pp 561-564, Oct. 2004.

[9] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, “Active shape models-their training and application”, *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38-59, Jan. 1995.