Static and Dynamic Lip Feature Analysis for Speaker Verification

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

As we all known, various speakers have their own talking styles. Hence, lip shape and its movement can be used as a new biometrics and infer the speaker's identity. Compared with the traditional biometrics such as human face and fingerprint, person verification based on the lip feature has the advantage of containing both static and dynamic information. Many researchers have demonstrated that incorporating dynamic information such as lip movement help improve the verification performance. However, which is more discriminative, the static features or the dynamic features remained unsolved. In this paper, the discriminative power analysis of the static and dynamic lip features is performed. For the static lip features, a new kind of feature representation including the geometric features, contour descriptors and texture features is proposed and the Gaussian Mixture Model (GMM) is employed as the classifier. For the dynamic features, Hidden Markov Model (HMM) is 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. Detailed evaluation for various static/dynamic lip feature representation is made along with a corresponding discussion on the discriminative ability. The experimental results disclose that the dynamic lip shape information and the static lip texture information contain much identity-relevant information.

Index Terms — lip feature, feature analysis, speaker verification

1. INTRODUCTION

Compared with the conventional accessing methods such as password or Personal Identity Number (PIN), the biometric features of a person can provide higher security level. The traditional biometric features such as fingerprint, iris, face, and hand have been proposed and used for person verification in many security systems. Visual information about the lip shape and movement has recently been used as a new biometric feature in a multimodal person verification system [1-4].

Recently, many researchers have proposed various kinds of feature representations to describe the lip information for visual speech recognition and visual speaker verification/authentication. Among them, image-based and model-based lip features are the most widely-used. Image-based features [5] are usually derived directly from the image after some form of filtering. Information lossless is the prominent advantage of this kind of features. However, they are usually of high dimension and high redundancy. In addition, it is also difficult for the classifier to extract speech relevant information from these features while ignoring the effects of scaling, rotation, translation and illumination. Model-based features [2,5,6,8] are of low dimensionality and invariant to the varying factors mentioned above. Moreover, some important speaker-relevant (speech-relevant) information can be easily derived from the lip model. Width and height of the lip are the most commonly-used model-based features for their easy derivation. Various kinds of lip contour descriptors such as the active shape model (ASM) vector [2], the geometric lip features [6], etc., are another kind of model-based features to describe the lip region. Recently, Matthews et al. [5] have proposed a new kind of model-based feature considering the intensity variation inside the outer lip contour (referred as the lip texture feature), which has shown effectiveness in visual speech recognition. Since the lip texture feature is usually of high dimension with much redundant information, PCA has been used for dimension-reduction in the literature [5].

All the model-based lip features used for visual speech recognition and visual speaker authentication/verification in the literature can be categorized into three kinds: geometric features, contour descriptors and texture features. In this paper, the derivation of all these three kinds of lip features is introduced. Moreover, a new kind of texture feature representation is proposed which may achieve better performance than the traditional PCA-based texture features.

As the lip feature is continuous in the time domain, the Hidden Markov Model (HMM) is usually adopted as the classifier for visual speaker verification and visual speech recognition since it performs well when dealing with the time-series data. Hence HMM is also employed in our approach as the classifier to evaluate the performance of the dynamic lip feature.

In summary, the experimental results reported in the literature [1-4, 6] have demonstrated that the lip region and

The work described in this paper is fully supported by NSFC Fund (60702043), the Shanghai Natural Science Fund (05ZR14080) and Sponsored by Shanghai Educational Development Foundation.

its movement contain rich identity-relevant information, which is very useful in visual speaker verification. However, one question is usually omitted in the literature: which is more discriminative, the static lip region or the dynamic lip movement? In this paper, a detailed analysis and discussion on the discriminative power of the static and dynamic lip feature is given. From the experimental results, it is shown that the static lip features have comparable discriminative power compared with the dynamic lip features especially when the entire corpus is limited.

The paper is organized as follows. In section 2, the lip modeling and lip feature extraction methods are introduced. Section 3 presents the classification techniques for both static and dynamic lip features. Detailed evaluation for discriminative power analysis along with the speaker verification results are also given in this section. Corresponding discussions on the evaluation results are given in the section 4 and section 5 draws the conclusion.

2. LIP MODELING AND FEATURE EXTRACTION

2.1 Lip contour extraction

In the previous work of our group [7,8], a lip segmentation technique [7] and a lip contour extraction method [8] have been proposed to derive the lip contour accurately and robustly from color lip images. The lip segmentation technique aims to partition all the pixels in the lip image into two parts: lip pixels and non-lip ones. Considering the presence of various kinds of disturbances such as beards, image noise and ambiguity, etc., a fuzzy-clustering based lip image segmentation process [7] is performed and then a membership map is generated which assigns a lip-class probability value to each pixel. With the membership map, a point-driven lip contour extraction method [8] is adopted to derive the lip contour in an efficient manner. Fig 1 has demonstrated some lip contour extraction results in a lip image sequence. Interested reader may refer to [7,8] for detailed introduction of the lip segmentation and contour extraction methods.

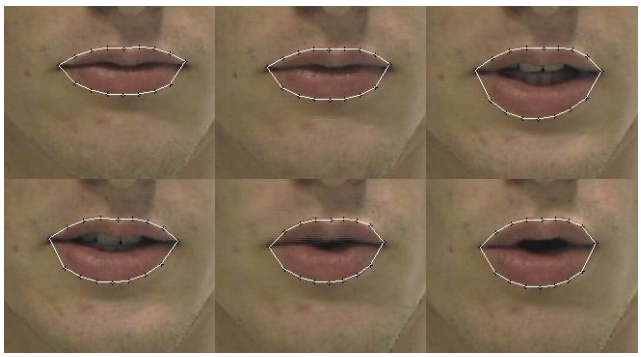

Fig. 1 Lip contour extraction results of several frames in a lip image sequence.

2.2 Lip feature extraction

In our evaluations, three kinds of lip features are extracted, including: i) the geometric features; ii) the contour descriptors; iii) the texture features. In the following subsections, the extraction method of the above three kinds of lip features is introduced.

2.2.1 Geometric lip features

The width and height of the lip are the widely-used geometric lip features in the literature. With the lip contour derived by [8], these geometric features can be obtained directly (as shown in Fig. 2).

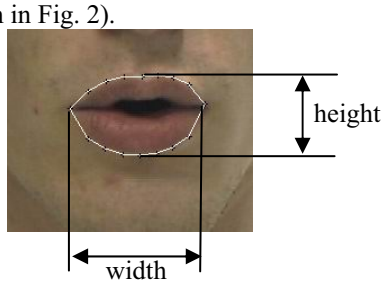

Fig.2 The geometric lip features

Note that due to the variations caused by different distances between the speaker and the camera, the geometric features are normalized against those of the first lip image in the sequence to avoid the above variations. Therefore, the geometric lip feature set contains: $\mathbf{f}_{\mathbf{geo}} = \{w_{normalized}, h_{normalized}\}$.

2.2.2 Contour descriptors

The geometric lip features alone cannot describe the lip region sufficiently since with similar width and height, the lip can have various kinds of shapes. The contour descriptor features aim to describe the lip shape information by a relatively low dimensional vector.

Due to the variation in translation, scaling and rotation, the coordinates of the contour-points cannot be used as the contour descriptors directly. The shape alignment scheme similar to that in ASM training [9] is adopted for contour feature normalization, which runs as follows:

1. Select a number of lip images (400 images in our experiment) with various kinds of lip shapes to build a training set. And extract the lip contour points for each lip image, which is represented by $\mathbf{x}_{\mathbf{s}} = \{x_1, y_1, x_2, y_2, ..., x_{16}, y_{16}\}$.
2. Align the lip contour points in the training set by the conventional iterative algorithm in [9] and calculate the average lip shape, $\overline{\mathbf{x}_{\mathbf{s}}}$, by averaging the aligned contour points.
3. The contour descriptor features can be derived by calculating the deviation between the align lip shape and the mean shape, i.e. $\mathbf{f}_{\mathbf{contour}} = \mathbf{x}_{\mathbf{s,aligned}} - \overline{\mathbf{x}_{\mathbf{s}}}$.

2.2.3 Texture lip features

Different from the geometric and contour descriptors, the texture lip feature aims to describe the intensity variations inside the lip region, which may contain the lip, teeth, tongue and oral cavity. Considering the variations caused by different lip shapes and lighting conditions, a two-stage normalization scheme is proposed as follows:

Shape Alignment: For each lip image being processed, the extracted contour points $\mathbf{x_s}$ and the mean lip shape $\overline{\mathbf{x_s}}$ are aligned and the entire region inside the outer lip contour is then projected onto the mean lip shape. Detailed shape alignment method is described in Appendix.

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 $\overline{\mathbf{x_s}}$, 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 $\{\mathbf{I_1}, \mathbf{I_2},\ldots,\mathbf{I_{400}}\}$.
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,
$$\mathbf{I_{nor,i}} = (\mathbf{I_i} - mean\_i \cdot \mathbf{1}) / amp \quad (1)$$
$$mean\_i = \mathbf{I_i} \cdot \mathbf{1} / m, \quad amp = \mathbf{I_i} \cdot \mathbf{I_{ref}} \quad (2)$$
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.,
$$\mathbf{I_{ref,new}} = \frac{1}{400}\sum_{i=1}^{400}\mathbf{I_{nor,i}} \quad (3)$$
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).

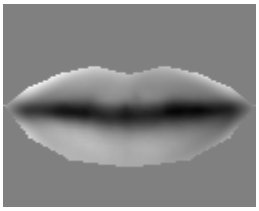

Fig.3 The reference intensity distribution map

2.3 Dimension reduction techniques by PCA and 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, two widely-used dimension reduction and feature extraction techniques, the Principle Component Analysis (PCA) and the Independent Component Analysis (ICA) [10], are employed to extract more robust features representing the lip region. For PCA, the low-dimension PCA features, $\mathbf{f_{PCA}}$, can be obtained by extracting the first several eigenvalues of the normalized contour or texture features [10]. To extract low-dimensional ICA feature, the original data is assumed to be a linear mixture of an unknown set of N statistically independent source shapes, i.e.,
$$\mathbf{f_{raw}} = a_1\mathbf{s_1} + a_2\mathbf{s_2} + \ldots + a_N\mathbf{s_N} = \mathbf{a_{data}}\mathbf{S_{data}} \quad (4)$$
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 [10] that maximizes the statistical independence between estimated sources is used. Hence, the ICA lip feature is represented by,
$$\mathbf{f_{ICA}} = \mathbf{a_{data}} = \mathbf{f_{raw}} \cdot \mathbf{S_{data}^T} \cdot (\mathbf{S_{data}^T}\mathbf{S_{data}})^{-1} \quad (5)$$

3. SPEAKER VERIFICATION BY STATIC AND DYNAMIC LIP FEATURES

3.1 Database and experimental setups

In order to evaluate the verification performance using various kinds of lip features introduced in section 2, a database consists of 40 speakers with 29 males and 11 females uttering the same phrase three-seven-two-five (3725) in English. Each speaker was asked to repeat the phrase for ten times and each utterance contains 90 lip images with size 110 by 90 lasting for 3 seconds.

3.2 Speaker verification by static lip features

In this evaluation stage, the lip information is taken as a static biometric feature like fingerprint or human face. In such case, the time information is omitted and the entire utterance can be regarded as a set of isolated lip images.

Training stage: the three lip image sequences (270 lip images) of each speaker are adopted to build the training data set. The Gaussian Mixture Model (GMM) is adopted as the classifier.

Verification stage: the remaining seven lip image sequences (630 lip images) of each speaker are adopted for testing. The test lip image is assigned to the speaker GMM with maximum likelihood and the average recognition accuracy among all the testing images (denoted by $R_{isolate}$) is adopted to evaluate the performance of the static feature set. With the assignment of every isolate image in the lip sequence, a voting process is carried out and the entire utterance is recognized as the speaker with maximum votes. Similar to that of the isolate lip image, the average recognition accuracy among all the testing sequences (denoted by $R_{sequence}$) is employed for evaluation.

Table 1 demonstrates the speaker verification performance by various static lip feature sets. As different pre-assigned training data samples may lead to different verification performance, one hundred random tests have been carried out and the average recognition value is employed for a more robust evaluation result. Note that the dimension of the contour/texture feature for PCA and ICA is set to (4,4) and (300,100) empirically which provides bettere verification result compared with other settings.

Feature Set	$R_{isolate}$ (%)	$R_{sequence}$ (%)
$\mathbf{f}_{geo}$	3.84	3.39
$\mathbf{f}_{contour,PCA}$	21.77	28.08
$\mathbf{f}_{contour,ICA}$	23.21	29.65
$\mathbf{f}_{texture,PCA}$	58.47	85.62
$\mathbf{f}_{texture,ICA}$	**60.17**	**87.12**
$\mathbf{f}_{geo}+\mathbf{f}_{contour,PCA}$	49.35	72.45
$\mathbf{f}_{geo}+\mathbf{f}_{contour,ICA}$	49.74	74.21
$\mathbf{f}_{geo}+\mathbf{f}_{texture,PCA}$	59.23	85.93
$\mathbf{f}_{geo}+\mathbf{f}_{texture,ICA}$	60.77	89.26
$\mathbf{f}_{geo}+\mathbf{f}_{contour,PCA}+\mathbf{f}_{texture,PCA}$	60.71	89.44
$\mathbf{f}_{geo}+\mathbf{f}_{contour,ICA}+\mathbf{f}_{texture,ICA}$	**62.21**	**90.67**

Table. 1 Speaker recognition accuracy in % by GMM with different kind of static feature sets.

From the table, the following issues can be observed: i) The geometric features alone are almost of no value for differentiating various speakers. The width and height of the lip usually changes much during the utterance and thus they only contain very limited identity-relevant information. ii) The lip contour information alone cannot provide satisfactory verification performance, either. However, taking both the geometric and contour features will greatly improve the performance which demonstrates that entire shape information can provide certain identity-relevant information. iii) Compared with PCA, ICA-based feature representations show better performance. In summary, speaker verification based on the static features alone can achieve satisfactory result (90.67% accuracy).

3.3 Speaker verification by dynamic lip features

In this evaluation stage, only the dynamic information of the lip (i.e. the first derivatives of the lip features, denoted by $\mathbf{f}'$) is employed to indicate the speaker's identity. A left to right, six states, continuous density Hidden Markov Model (HMM) with diagonal covariance matrix Gaussian model associated with each state is adopted as the classifier.

Training stage: three utterances of each speaker are adopted to build training data set and the Baum-Welch algorithm following the Maximum Likelihood (ML) criterion has been used for training the HMM.

Verification stage: the remaining seven utterances are used as testing data. The Viterbi algorithm for recognition and the testing lip sequence is assigned to the speaker model with maximum likelihood. The average recognition rate (denoted by $R_{dynamic}$) is adopted to evaluate the verification performance. Similar to that of the static features, one hundred random tests are performed and the average recognition performance is listed in Table 2.

Feature Set	$R_{dynamic}$ (%)
$\mathbf{f}'_{geo}$	50.32
$\mathbf{f}'_{contour,\ PCA}$	51.55
$\mathbf{f}'_{contour,\ ICA}$	53.48
$\mathbf{f}'_{texture,\ PCA}$	81.27
$\mathbf{f}'_{texture,\ ICA}$	83.98
$\mathbf{f}'_{geo}+\mathbf{f}'_{contour,\ PCA}$	73.55
$\mathbf{f}'_{geo}+\mathbf{f}'_{contour,\ ICA}$	73.48
$\mathbf{f}'_{geo}+\mathbf{f}'_{texture,\ PCA}$	81.42
$\mathbf{f}'_{geo}+\mathbf{f}'_{texture,\ ICA}$	84.52
$\mathbf{f}'_{geo}+\mathbf{f}'_{contour,\ PCA}+\mathbf{f}'_{texture,\ PCA}$	83.23
$\mathbf{f}'_{geo}+\mathbf{f}'_{contour,\ ICA}+\mathbf{f}'_{texture,\ ICA}$	**84.78**

Table. 2 Speaker recognition accuracy in % by HMM with different kind of dynamic feature sets.

From Table 2, similar observations can be made: i) the discriminative power of the dynamic texture feature is much higher than that of the geometric features and contour descriptors. ii) The dynamic geometric features could be useful to improve the verification performance for the dynamic contour descriptors while could only contribute little for the dynamic texture features. iii) ICA-based features also perform better than the PCA-based features in the above dynamic feature evaluation. As a result, speaker verification by the dynamic features alone can achieve the accuracy of 84.78%.

4. DISCUSSIONS

Recent research discloses that lip shape and movement contain useful information for visual speech recognition and visual speaker verification. Comparing with the static lip information (lip shape, intensity, etc.), dynamic lip motion contains more speech-relevant information since human speech is a continuous process in nature. However, for differentiating different speakers, the discriminative ability of the static and dynamic lip information has not been addressed much in the literature.

From the experimental results given in Section 3, the following can be concluded:

i) For the geometric features and contour descriptors, the dynamic representation achieves much better performance compared with the static representation, which demonstrates that the speaker's identity information is contained in the lip shape dynamics rather than any static lip shape.

ii) For the texture lip features, although the verification result from any isolate lip image is not so accurate (62.21%), the overall recognition performance after voting is more reliable (above 90%) compared with that of the dynamic texture features (83.98%), which indicates that the speaker's identity information is more likely to be contained in the static lip texture rather than the texture changes. In addition, the lip texture feature is shown to have the highest discriminative power among all the lip features.

5. CONCLUSIONS

This paper proposes a new lip feature representation for speaker identity authentication, which contains the geometric features, contour descriptors and texture-based features. Based on the above lip features and proper classifiers, i.e. GMM for static features and HMM for dynamic features, a detailed evaluation to analyze the discriminative ability of the static and dynamic lip information is performed. The experimental results have demonstrated that: i) the texture information contains much identity-related information compared with the shape and contour information; ii) for the contour descriptors and texture features, the ICA feature extraction method is more appropriate than the conventional PCA; iii) for the lip shape dynamics contain more information than the static lip shape while the static lip texture information has more identity-relevant information than the texture changes.

6. APPENDIX

Two lip shapes $\mathbf{x}_s$ and $\overline{\mathbf{x}_S}$ are aligned by minimizing,

$$E = \left(\overline{\mathbf{x}_s} - M(s,\theta)\mathbf{x}_s - \mathbf{t}\right)^T \left(\overline{\mathbf{x}_s} - M(s,\theta)\mathbf{x}_s - \mathbf{t}\right) \qquad \text{(A.1)}$$

where $M(s,\theta)$ is the affine transformation of $\mathbf{x}_s$ [6],

$$M(s,\theta)\begin{bmatrix} x \\ y \end{bmatrix} = \begin{pmatrix} s\cos\theta \cdot x - s\sin\theta \cdot y \\ s\sin\theta \cdot x + s\cos\theta \cdot y \end{pmatrix} \qquad \text{(A.2)}$$

By minimizing the cost function E, the scaling factor *s*, rotation angle θ and the translation **t** are recorded for subsequent processing.

7. REFERENCES

[1] A. Kanak, E. Erzin, Y. Yemez, A.M. Tekalp, "Joint audio-video processing for biometric speaker identification", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong SAR, China, vol. 3, pp. 561-564, July 2003.

[2] J. Luettin, N. A. Thacker and S. W. Beet, "Speaker identification by lipreading", Proceedings of Fourth International Conference on Spoken Language (ICSLP '96), Philadelphia, PA, USA, vol.1, pp. 62-65, 1996.

[3] 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.

[4] G.B. Ou, X. Li, X.C. Yao, H.B. Jia, Y.L. Murphey, "Speaker identification using speech and lip features", Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, vol.4, pp. 2565-2570, July, 2005.

[5] 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.

[6] 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.

[7] S. L. Wang and A.W.C. Liew, "Robust Lip Region Segmentation for Lip Images with Complex Background", *Pattern Recognition* (40), no. 12, pp. 3481-3491, Dec. 2007.

[8] 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.

[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.

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