LIP FEATURE EXTRACTION BASED ON AUDIO-VISUAL CORRELATION

M. E. Sargın, E. Erzin, Y. Yemez, A. M. Tekalp

Multimedia Vision and Graphics Lab., College of Engineering,
Koç University, Sarıyer, Istanbul, 34450 Turkey
msargin,eerzin,yyemez,mtekalp@ku.edu.tr

ABSTRACT

In this paper, the lip feature that has the highest correlation with audio features is investigated. Audio features are selected as Mel Frequency Cepstral Coefficients (MFCC) of the audio signal. Three different lip features are considered for the visual lip information, where these features are 2D DCT coefficients of the intensity based image and the optical flow vectors within the lip region, and the distances between pre-defined points on the lip contour which carries the lip shape information. In this study, we present two techniques based on class conditional probability analysis and canonical correlation analysis to estimate and compare the correlations between audio feature and each lip feature. The lip feature, which has the highest correlation to audio features, is identified among the above lip features. Isolation of lip features, which are highly correlated with audio signal, can be used for audio-visual speech recognition, audio-visual lip synchronization and estimation of lip shapes using audio signal for visual synthesis.

1. INTRODUCTION

Multimodal speech and speaker recognition systems use the visual and auditory information together in order to increase the recognition performance [1, 2, 3, 4]. The performance of the multimodal recognition systems increases when the correlation between visual and auditory information increases. Since it has been a widely accepted fact that lip-motion is highly correlated with speech, it is a common practice to use lip-motion as visual feature.

In literature, several lip features are used in multimodal recognition systems. Most common lip features are, intensity image of lip frame, optical flow vectors calculated for each lip frame and lip-shape parameters. The use of intensity image of lip frame has the disadvantage that it is variant to lighting conditions. Optical flow vectors are invariant to lightning conditions whereas they often introduce error due to occlusion in the interior lip region. Lip shape parameters seems to be robust to former effects if outer and inner lip contours are extracted exactly. However automatic tracking algorithms[5] fail under poor light and extraordinary lip shapes. In those cases, manual selection can be a possible solution however it is not feasible for realtime systems.

Although the performances of these different multimodal recognition systems are extensively studied individually, there is relatively little work on finding the relative performances of different lip features in recognition systems, based on correlation analysis. In this paper, correlation analysis of three different lip features with auditory feature is studied. The most correlated lip feature to speech signal is determined using two methods. In class conditional probability analysis the feature that has the highest class-conditional pdf match is investigated. In second method, the feature that has the highest canonical correlation coefficients is investigated.

THIS WORK HAS BEEN SUPPORTED BY EUROPEAN UNION 6TH FRAMEWORK PROGRAM UNDER THE PROJECT (SIMILAR) FP6-507609 (HTTP://WWW.SIMILAR.CC).

2. AUDIO-VISUAL FEATURE EXTRACTION

2.1 Extraction of Speech Features

In this study, speech signal is represented with 13 Mel Frequency Cepstrum Coefficients (MFCC) including the energy term. MFCC coefficients are calculated over 25 ms windows for each 10 ms frame, where the resulting speech feature rate is 100 fps.

2.2 Extraction of Lip Features

Prior to lip feature extraction global head motion compensation is applied to face images to avoid noise and undesired components in the lip features. In the global head motion compensation, each face image frame is aligned using 2D parametric motion estimator. For every two consecutive face images 12 quadratic motion model parameters are calculated using the method described in [9]. Corresponding 12 parameter motion model relates each pixel (x_i, y_i) at frame i with pixel (x_{i+1}, y_{i+1}) at frame $i+1$ with the equation

$$\begin{bmatrix} x_{i+1}-x_i \\ y_{i+1}-y_i \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} a_3 & a_4 \\ a_5 & a_6 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} a_7 & a_8 & a_9 \\ a_{10} & a_{11} & a_{12} \end{bmatrix} \begin{bmatrix} x_i^2 \\ x_i y_i \\ y_i^2 \end{bmatrix} . \quad (1)$$

Then each face image $i+1$ is warped according to 12 motion parameters calculated between frames i and $i+1$.

The visual information is represented as lip shape and lip motion with three different features. First lip feature is the 2 dimensional discrete cosine transform (2D DCT) of the optical flow vectors. The optical flow vectors are extracted from the lip frames using hierarchical Lucas-Kanade technique. The first 25 baseband 2D DCT zig-zag scan coefficients are used for each horizontal and vertical motion vectors. Second lip feature is the intensity based 2D DCT of the lip frames. Likewise optical flow vectors, the first 25 baseband 2D DCT zig-zag scan coefficients are used for each axis. Third and last, lip features are selected as the lip-shapes. The lip-shape feature vectors are composed of 8 parameters where each parameter is the distance between key points on lip contour representing the horizontal and the vertical opening of lip. The key points and sample distances can be seen in Figure 1. Lip contour and key points are extracted from lip frames using the algorithm described in [5]. This technique consist of extraction of key points followed by fitting optimal four third degree polynomials and two line segments to 5 key points and three key points that form Cupidon's bow respectively. These visual features were also used in our previous speaker identification system [6, 7].

Figure 1: Lip Shape Parameters

2.3 Synchronization of Audio-Visual Features

As mentioned in 2.1, auditory features are the MFCC coefficients which have the frame rate of 100 fps whereas all visual features have frame rate of 15 fps. In order to synchronize audio and visual features, visual features are interpolated to 100 fps rate. For intensity based 2D DCT of lip frames and lip shape parameters linear interpolation is used. However, for the interpolation of 2D DCT of optical flow vectors, the optical flow vectors are first interpolated, and then the 2D DCT coefficients of interpolated optical flow vectors are calculated. Linear interpolation of intensity based 2D DCT features and lip shape parameter features are calculated as in [6],

$$f_a^k = f_a\left(\frac{k}{100}\right), \quad k = 0,1,2,3\cdots, \tag{2}$$

$$f_l^i = f_l\left(\frac{i}{15}\right), \quad i = 0,1,2,3\cdots, \tag{3}$$

$$\hat{f}_l^k = (1-\alpha^k) f_l^{i^*} + \alpha^k f_l^{i^*+1} \tag{4}$$

where the audio and the visual features are represented at time instants $k\frac{1}{100}$ and $i\frac{1}{15}$ seconds, respectively as f_a and f_l, and the interpolated lip features $\hat{f}_l$ are extracted with $i^* = \lfloor \frac{3k}{20} \rfloor$ and $\alpha^k = \frac{3k}{20} - i^*$.

Since optical flow vectors carry motion information between two consecutive lip frames, interpolation of these motion vectors need to be performed correspondingly. Let, l_1 and l_2, be two consecutive lip frames and f_1 is the associated motion vector of l_1 indicating where each pixel in l_1 will move in l_2. Suppose frame l_i, where $1 \leq i \leq 2$, will be interpolated and the optical flow vectors for the new frame sampling will be calculated. In this scenario, let the optical flow vector at pixel (x,y) of l_1 be (a,b). Then, the optical flow vectors after an interpolation, which assumes uniform motion, can be calculated as $(\frac{a}{2},\frac{b}{2})$ at pixel (x,y) of l_1, and $(\frac{a}{2},\frac{b}{2})$ at pixel $(x+\frac{a}{2}, y+\frac{b}{2})$ of $l_{1.5}$. The proposed motion interpolation scheme is presented in Figure 2. However, the optical flow vectors from l_1 to l_2 do not intersect all the pixels of l_i. Hence, an unassigned optic flow vector at a pixel of l_i is assigned to be the average of the optic flow vectors at neighboring pixels.

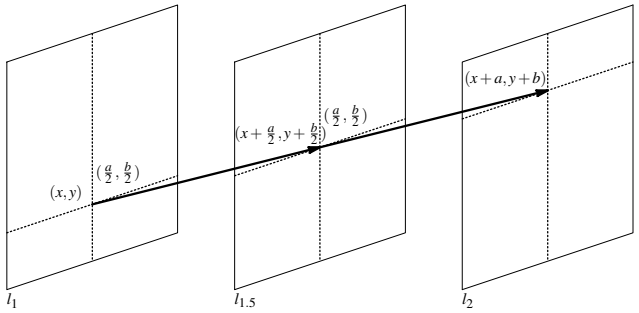

Figure 2: Motion interpolation scheme.

3. AUDIO-VISUAL CORRELATION ANALYSIS

3.1 Cross Correlation Matrix

The cross correlation matrix is examined to observe the correlation of audio-visual features. After synchronization, audio-visual features of the n-th frame are concatenated into a single feature vector V_n, and the cross correlation matrix C of V_n is calculated. Let the dimension of MFCC coefficients be D_a and the dimension of visual features be D_l, then the cross correlation matrix C is $(D_a + D_l) \times (D_a + D_l)$. The non-diagonal terms of C will be the correlation coefficient between each element of auditory and visual features. The cross correlation matrices for three words spoken by same person are given in Figure 3. The first and the second correlation matrix visualizations correspond to same word uttered. However, the third correlation matrix visualization corresponds to a different word uttered by the same speaker. In these visualizations, the vertical axis of the cross correlation matrix is the auditory features, and the horizontal axis is the first 40 zig-zag scan of the 2D DCT coefficients of optical flow vectors for horizontal and vertical motion vectors. Note that, patterns in the first and the second matrices are similar to each other, whereas patterns in the third matrix is different. This is expected since correlations between the audio-visual feature coefficients vary with the speech content.

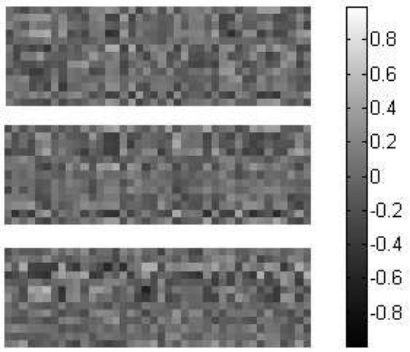

Figure 3: Cross Correlation Matrices

3.2 Canonical Correlation Analysis

Canonical Correlation Analysis CCA , which is a way of measuring the linear relationship between two multidimensional variables, was first introduced by H. Hotelling [10]. CCA searches for two sets of basis vectors related with each variable, so that the correlation of variables in new basis is diagonal and the diagonal elements are maximized. An important property of canonical correlations is that they are invariant with respect to affine transformations of the variables. This is the most important difference between CCA and ordinary correlation analysis which highly depend on the basis in which the variables are described.

Let two multidimensional biometric signals are represented with $\mathbf{x}$ and $\mathbf{y}$. Further let the projection matrices be $\mathbf{w}_x$ and $\mathbf{w}_y$ such that the correlations between the projections of $\mathbf{x}$ and $\mathbf{y}$ onto $R(\mathbf{w}_x)$ and $R(\mathbf{w}_y)$ are mutually maximized. The problem becomes maximization of ρ

$$\rho = \frac{E[\acute{\mathbf{x}}\acute{\mathbf{y}}^T]}{\sqrt{E[\acute{\mathbf{x}}\acute{\mathbf{x}}^T]E[\acute{\mathbf{y}}\acute{\mathbf{y}}^T]}} \tag{5}$$

where $\acute{\mathbf{x}} = \mathbf{w}_x^T\mathbf{x}$ and $\acute{\mathbf{y}} = \mathbf{w}_y^T\mathbf{y}$, Let $\mathbf{x}$ and $\mathbf{y}$ be zero mean random variables. The total covariance matrix is defined as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = E\left[\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \begin{bmatrix} \mathbf{x}^T & \mathbf{y}^T \end{bmatrix} \right] \tag{6}$$

Where, $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are within set covariance matrices, and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$ is between-set correlation matrix. The canonical correlations between $\mathbf{x}$ and $\mathbf{y}$ can be found by solving the eigenvalue equations

$$\begin{aligned} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x &= \rho^2\mathbf{w}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{w}_y &= \rho^2\mathbf{w}_y \end{aligned} \tag{7}$$

where the eigenvalues ρ^2 are the squared canonical correlations and the eigenvectors $\mathbf{w}_x$ and $\mathbf{w}_y$ are the normalized canonical correlation basis vectors. Only one of the eigenvalue equations needs to be

solved since the solutions are related by

$$\begin{array}{l} \mathbf{C}_{xy}\mathbf{w}_y = \rho\lambda_x\mathbf{C}_{xx}\mathbf{w}_x \\ \mathbf{C}_{yx}\mathbf{w}_x = \rho\lambda_y\mathbf{C}_{yy}\mathbf{w}_y \end{array} . \tag{8}$$

Where,

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\mathbf{w}_y^T\mathbf{C}_{yy}\mathbf{w}_y}{\mathbf{w}_x^T\mathbf{C}_{xx}\mathbf{w}_x}}. \tag{9}$$

In this study, CCA is applied between proposed lip features and audio features. For different shift sizes $\|\rho\|_2^2$'s between circularly shifted lip features and audio features is calculated. Maximum $\|\rho\|_2^2$ is achieved when two sequences have highest correlation. This technique can be applied to lip synchronization problem by using audio-visual correlation.

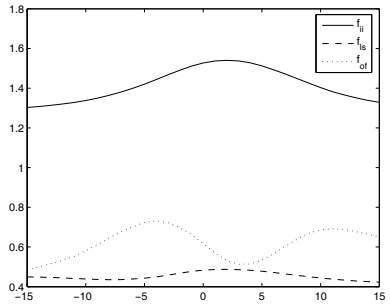

Figure 4: $\|\rho\|_2^2$ for Different Shift Sizes

3.3 Statistical Modeling of Audio-Visual Correlations

Statistical modeling of audio-visual correlations can be performed by fitting audio-visual joint probability density function (pdf), $p(f_a, f_l)$, over some training data, and the fit of the joint pdf can be tested on an independent test data. However, training of joint pdf is not practical. Hence, in this study audio-visual correlations are modeled by fitting class conditional pdf, $p(f_a|f_l)$, where for a given lip feature class the pdf of audio features are modeled. The determination of lip feature classes are done using vector quantization over a predefined training database. Let the vector quantized lip features are represented by $\{\hat{f}_l\}$. Having temporally synchronized audio-visual features, one can extract all the audio features that corresponds to each quantized lip feature cluster. Let us represent the collection of such audio features for each lip feature class as,

$$S_{\hat{f}_l^n} = \{f_a^n(k)|\hat{f}_l^n = VQ(f_l(k)), \quad \forall k\}, \quad n = 1,2,...,N \tag{10}$$

where k is the frame index and N is the number of clusters in the lip feature vector quantization. Statistical modeling for each such collection of audio features, $S_{\hat{f}_l^n}$, is done using Gaussian Mixture Models (GMM), where each class conditional pdf is estimated as a weighted sum of Gaussian distributions,

$$p(f_a|\hat{f}_l^n) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\mu_m, \Sigma_m), \tag{11}$$

where M is the number of Gaussian mixtures in the GMM model. The conditional pdf estimation avails us to model correlations that we observed in the cross-correlation matrix visualization in Figure 3.

3.3.1 Statistical Model Verification

After estimating the probabilistic correlation model, the goodness of the statistical model fitting can be tested for verification and lip feature selection purposes. In the verification phase, an audio-visual database, which is independent from the pdf training database should be used. A goodness of fitting probability measure can be defined over such a verification database,

$$\rho_l = \prod_{f_a \ \mathtt{st.}\ f_a \in S_{\hat{f}_{l_i}}} p(f_a|\hat{f}_l). \tag{12}$$

The goodness measure ρ_l represents the probability of observing all audio features synchronous to lip features f_l using the class conditional pdf models. The goodness measure ρ_l can be calculated for each lip feature, which are 2D DCT of optical flow (f_{of}), 2D DCT of intensity image (f_{ii}), and lip shape parameters (f_{ls}). The lip feature, which maximizes the goodness measure ρ_l over all lip features,

$$l^* = \arg\max_{l=f_{of},f_{ii},f_{ls}} \rho_l \tag{13}$$

is selected as the most correlated lip feature to audio signal. The main idea behind this fact is that the bigger ρ_l is, the better class conditional pdf estimated using training data fits test data. In other words, if ρ_l is big we have a better estimation of the class conditional pdf of the test data by using training data. If two features are highly correlated we could estimate the class conditional pdf of the test data with small estimation error or equivalently big ρ_l.

3.3.2 Estimation of Lip Features

A possible application of the class conditional probability modeling of audio-visual features is to estimate the corresponding lip feature $\tilde{f}_l$ given the audio feature f_a. Such a probabilistic mapping is possible by maximizing class conditional pdf over all quantized clusters of lip features,

$$\tilde{f}_l = \arg\max_{f \in VQ(f_l)} p(f_a|f). \tag{14}$$

The estimated lip feature corresponding to an audio frame/feature is assigned as the best VQ centroid that maximizes the above class conditional pdf. This estimation can be compared with the true quantized lip feature. Since the temporal correlations are not employed in this estimation, one should not expect high identification rates. However, the number of true estimations can be another lip feature selection criteria that emphasize audio-visual correlations. Selected results on the lip feature estimation are presented in Section 4.

3.3.3 Estimation of Audio Features

Alternatively, one can interchange audio and lip features in the proposed statistical model, that is, the class conditional pdf of lip features given vector quantized audio features can be estimated. In this new framework, the audio features can be estimated by maximizing class conditional pdf as,

$$\tilde{f}_a^l = \arg\max_{f \in VQ(f_a)} p(f_l|f). \tag{15}$$

Once the estimate of the audio feature, $\tilde{f}_a^l$, is found, the closeness of the estimate to the true audio feature can be calculated by Euclidian distance. In general, the expected value of the distance between estimated audio feature and the true audio feature,

$$D_l = E\{||\tilde{f}_a^l - f_a||\}, \tag{16}$$

is another performance measure for the audio-visual correlations that we modeled using class conditional pdfs. As similarly defined in section 3.3.1, the lip feature f_l, which minimize the expected distortion D_l, can be set as the most correlated lip feature with audio. Results on the audio feature estimation are presented in Section 4.

Lip Feat. (l)	VQ	$log(\rho_l)$	Corr. Est.
f_{ii}	64	-827398.528656	848
	32	-631799.067095	1586
	16	-515752.398466	2398
f_{ls}	64	-453659.261208	1553
	32	-420744.753509	2217
	16	-400447.717235	3869
f_{of}	64	-437877.491851	1555
	32	-399796.107598	2372
	16	-387340.167759	3590

Table 1: Relative Performance Measures

Lip Feat. (l)	D_l
f_{ii}	24.73
f_{ls}	24.70
f_{of}	24.47

Table 2: Auditory Feature Estimation Performance Measures for 32 Level VQ

4. RESULTS

The experiments of the proposed technique are conducted by using a portion of the MVGL-AVD database [8]. In our experiments 30 subjects are used where each subject utters ten repetitions of her/his name. The class conditional densities $p(f_a|\hat{f}_l)$ are modeled by using the first five utterances of all subjects. The last five utterances of all subjects are used in testing and verification of the probabilistic model. Relative performance measures for different number of quantization levels are presented in Table 1. Relative auditory feature estimation performances for 32 level VQ are presented in Table 2. Note that, in this case only 32 level quantization is applied by considering the number of different phonemes in the database. In Table 1, the correct estimation column presents the number of frames that are estimated correctly among the whole test data, which includes 20428 frames.

The results of CCA showed that 2D DCT of intensity based images has the highest correlation. This result achieved if we represent audio-visual features in the basis $\mathbf{w}_a$ and $\mathbf{w}_v$ respectively by using the formulation given in Section 3.2.

Table 1 clearly presents that, the 2D DCT of optical flow vectors for 64, 32 and 16 level VQ extracts the largest ρ_l probabilities. This result is also supported in Table 2, where again the 2D DCT of optical flow vectors attain the minimum expected Euclidian distance among all lip features. A similar and consistent conclusion is also observed by investigating the correct estimation rates that are presented in Table 1. However, only in 16 level VQ correct estimation rate for 2D DCT of optical flow vectors is lower than lip shape parameters. This could be as a result of 16 level VQ, which could not able to represent all visemes in the database. Hence, one can conclude that the 2D DCT of optical flow vectors has the highest correlation with audio among the three lip features that are considered in this study.

5. CONCLUSIONS

In this study, the lip feature that has the highest correlation with audio features is investigated. Among the three different lip features, 2D DCT coefficients of the intensity based image and the optical flow vectors, and the lip shape information, the 2D DCT of optical flow vectors are found to have the highest correlation with audio. However CCA showed that a new set of basis vectors can be found for 2D DCT of intensity based image such that it has the highest correlation with audio features. Although two signals have highest correlation, using new basis vectors, as in the case of PCA or LDA, may decrease overall system performance. As a future work, overall system performance of using new features extracted from originals, such that audio-visual correlation is maximized, will be examined. Proposed methodology can be used as feature extractor instead of PCA or LDA for feature level fusion.

Isolation of these lip features, which carry highest audio-visual correlation, can further be used for audio-visual speech recognition, audio-visual lip synchronization and estimation of lip shapes using audio signal for visual synthesis.

REFERENCES

[1] Potamianos, G.; Neti, C.; Gravier, G.; Garg, A.; Senior, A.W., Recent advances in the automatic recognition of audiovisual speech, Proceedings of the IEEE, Vol.91, Iss.9, Sept. 2003 pp. 1306- 1326

[2] Potamianos, G.; Neti, C.; Huang, J.; Connell, J.H.; Chu, S.; Libal, V.; Marcheret, E.; Haas, N.; Jiang, J., Towards practical deployment of audio-visual speech recognition, Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on, Vol.3, Iss., 17-21 May 2004 pp. iii- 777-80 vol.3

[3] Gurbuz, S.; Tufekci, Z.; Patterson, E.; Gowdy, J.N., Independent information from visual features for multimodal speech recognition, SoutheastCon 2001. Proceedings. IEEE, Vol., Iss., 2001 sa. 221-228

[4] Luhong Liang; Xiaoxing Liu; Yibao Zhao; Xiaobo Pi; Nefian, A.V., Speaker independent audio-visual continuous speech recognition, Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on, Vol.2, Iss., 2002 pp. 25- 28 vol.2

[5] N. Eveno, A. Caplier, P.-Y. Coulon, "Accurate and Quasi-Automatic Lip Tracking," IEEE Trans. on Circuits and Systems for Video Technology vol. 14, Iss. 5, pp.706-715, May 2004.

[6] A. Kanak, E. Erzin, Y. Yemez and A. M. Tekalp, "Speaker Identification Using Multimodal Audio- Video Processing," IEEE Int. Conf. on Image Processing, Barcelona, Spain, September 2003.

[7] H.E. Cetingul, Y. Yemez, E. Erzin, A.M. Tekalp "Discriminative Lip-Motion Features for Biometric Speaker Identification," IEEE Int. Conf. on Image Processing, Singapore, 2004.

[8] E. Erzin, Y. Yemez, A. M. Tekalp, "Multimodal Speaker Identification Using an Adaptive Classifier Cascade based on Modality Reliability," accepted for publication on IEEE Transactions on Multimedia, March 2004.

[9] J.M. Odobez, P. Bouthemy, Robust multiresolution estimation of parametric motion models. Journal of Visual Communication and Image Representation, 6(4):348-365, December 1995.

[10] H. Hotelling. Relations between two sets of variates. Biometrika, 28:321 377, 1936.