Feature Extraction in Speechreading

Jun He
information college, NanChang university
Email: boxhejun@tom.com

Hua Zhang
NanChang University
Email : zhanghua_lab@163.com

***Abstract*—To solve the problem of feature extraction in speechreading, several appearance-based feature extraction method are compared and a new improved LDA algorithm is proposed in this paper. In speech or speechreading recognition application, Linear Discriminant Analysis(LDA) usually choose syllable、HMM state or other units as class unit. but the feature dimensionality reduction direction based on this traditional LDA have no direct relations with recognition accuracy,To this problem, A LDA algorithm based on Object (LDAO) which is fit for isolated words recognition in speechreading is proposed, LDAO choose the objects to be recognized as class unit to Linear Discriminant Analysis, which guarantees feature extraction follow the most discriminant directions among objects in theory. Subsequently, training and recognizing method for LDAO was also given. All experiments were performed on bimodal database, Experimental results showed that this algorithm is superior to any other appearance-based feature extraction algorithm in speechreading. Specifically, LDAO is better than DCT+LDA about 3%.**

***Index Terms*—speechreading, feature extraction, LDA, LDAO**

I. INTRODUCTION

Speechreading, which gets visual speech information by exploiting the visual modality of the speaker's mouth region in addition to the traditional audio modality, aims at improving speech recognition under noisy environment. Feature extraction is one of the most challenging part in speechreading. In general, feature extraction methods contain three categories[1]: the first is shape-based feature extraction method, which assumes that most speech information is contained in the contours of the speaker's lips, or more generally in the face contours, e.g. lip, jaw and cheek. Geometric-type feature was widely used in early time, such as mouth width, height,and area[2][3][4][5][6] , Fourier and image moment descriptors of the lip contours[7], later on , extracting the parameters of lip'shape model was paid more attention[8], such as active shape model(ASM)[9]; the second is appearance-based method. In contrast, appearance-based method assume that all pixels in ROI are informative speech features[10][11][12][13][14][15][16].To reduce high dimensionality of feature vector, some linear transformation are introduced, such as Principal Component Analysis (PCA) [17], Discrete Cosine Transformation (DCT)[18], LDA[23][25], Discrete Wavelet Transformation (DWT), etc; the third is hybrid method, which concatenate the former two features into a joint shape and appearance feature vector, e.g. active appearance model (AAM)[19[20][21]] is learned on such vectors.

Shape-based feature often has low feature dimensionality and apprehensible feature concept. But it lose too much speech information, so speechreading recognition accuracy is often rather low based on this feature[1]. Hybrid method has advantage in conception, but training model takes too long time, so it doesn't fit for real-time application; Appearance-based feature make use of all pixels information in Region Of Interest (ROI) , speech information is kept perfectly, But it's sensitive to scale,rotation, illumination and often has very high feature dimensions.

Johns Hopkins Summer 2000 Workshop on audio-visual automatic speech recognition (ASR) in the large-vocabulary, continuous speech domain performed a comparision among above three feature extraction methods on the same database[22], experiments showed that appearance-based method achieved best performance: Error accuracy is 58.14%, meanwhile, AAM algorithm error accuracy is 65.66%, Figure 1 shows the appearance-based DCT+LDA feature extraction method by Potamianos. which contain three linear transformation. firstly, ROI was performed DCT and 24 dominating DCT coefficients were retained; secondly, mean subtraction normalization was performed, and a three-dimensional ROI containing a sequence of adjacent 15 frames was built as super-frame feature vector, then LDA choosing HMM state as class unit was performed on these data; at last, Maximum likelihood linear regression (MLLR) was performed to carry out speaker adaptation. The key of this feature extraction method is LDA which has been successfully used in speech recognition. Haeb-Umbach first introduced LDA to speech recognition successfully[23], the rationale is connecting some adjacent speech feature and build a new feature vector in advance, then perform LDA based on subphoneme of German. Subsequently, different unit was used as class unit in LDA, such as phoneme, syllable, subsyllable or HMM state etc[24]. Because of the excellent segmenting data capability of Viterbi algorithm,

Manuscript received July 1, 2009; revised August 20, 2009; accepted September 25, 2009.

doi:10.4304/jsw.5.7.705-712

HMM state was recognized as the best class unit of LDA[25].

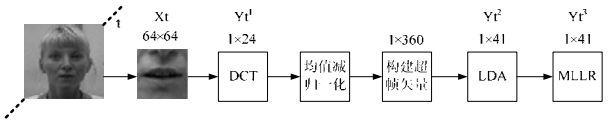

Figure 1. Feature extraction process based on DCT+LDA

However, no matter what kind of class units LDA choose, the obtained LDA projection direction is not associated with word recognition accuracy directly, which means the recognition accuracy based on these features extraction direction is not optimal. To this question, An improved LDA algorithm LDAO is proposed in this paper, the rationale is choosing objects to be recognized as class unit to LDA, the obtained projection direction from LDA denotes the optimal direction for object recognition. So in theory the LDAO algorithm is optimal. Experiments on speaker-dependent bimodal database showed that this algorithm is superior to any other appearance-based feature extraction methods.

In this paper, section 2 make a comparison between several traditional appearance-based feature extraction methods; section 3 explain the rationale of LDAO algorithm in detail; section 4 propose a method for LDAO training; in section 5 we compare LDAO with other appearance-based algorithms; Finally, section 6 concludes this paper with a summary and a brief discussion.

II. APPEARANCE-BASE EXTRACTION ALGORITHM

A. Feature extraction based on PCA

PCA aims at getting the dominating axis of the biggest variance of all the samples. Feature dimensionality reduction based on PCA is projecting original feature to the obtained subspace. It's the optimal data compression method between the reconstructed matrix and original matrix in the meaning of Minimum Mean Squared Error(MMSE). PCA is based on KL transformation in mathematical calculation, For samples $\{x_t\}$ t=1,2…,T, Feature extraction based on PCA is depicted as follows:

1.calculate the mean of all samples

$$\mu = \sum_{i=1}^{T} x_t \tag{1}$$

2. calculate covariance matrix

$$R = \frac{1}{N}\sum_{t=1}^{T}(x_t - \mu)(x_t - \mu)^{'} \tag{2}$$

3.compute the eigenvalue λ_1 , λ_2 ,…, λ_N and eigenvector $q_1, q_2, ..., q_N$ of covariance matrix, rank the eigenvalue according $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$, then use those eigenvector corresponding to the former M bigger eigenvalue build transformation matrix A =(q_1 , q_2 ,…, q_M).

For a new sample x, dimensionality reduction can be executed on equation(3), actually, feature extraction is project sample x to the subspace constructed by these eigenvectors axis. it's interesting that each eigenvector can be reshaped and form a picture which is very like lip. And often these pictures is called eigenlip. In figure 2. eigenlip is derived from the former 3 dominating eigenvector axis. The last picture is the mean lip of all samples.

$$y = A^T(x - \mu) \tag{3}$$

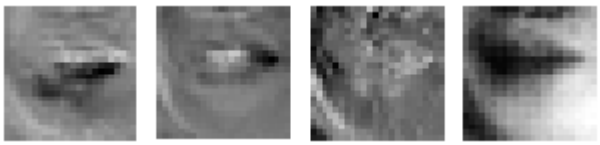

Figure2. The left three pictures are eigenlip corresponding to the former3 dominating eigenvectors. The last picture is the mean lip of all samples.

B. Feature extraction based on DCT

It has been demonstrated in experiment that DCT performs a little better than DWT and WAL[1]. Compared with PCA, Although DCT is suboptimal at data compression, DCT has two advantages: one is that DCT is performed on the picture itself, unlike PCA must depend on all samples; the other is DCT has fast algorithm similar to Fast Fourier Transform (FFT).

2D-DCT can save image information with only a little coefficients. so is often used to realize lossy image compression. DCT can be expressed as the sum of cosine function with different amplitude and frequency. For an image x(m,n), 2D-DCT is defined as follows:

$$y(u,v) = a(u)a(v)\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} x(m,n)\cos\left[\frac{(2m+1)u\pi}{2M}\right]\cos\left[\frac{(2n+1)v\pi}{2N}\right] \tag{4}$$

Where $y(u,v)$ is the DCT coefficient of X ,a(u),a(v) is:

$$a(u) = \begin{cases} \sqrt{\frac{1}{M}}, & k = 0 \\ \sqrt{\frac{2}{M}}, & k = 1,2,...M-1 \end{cases} \qquad a(v) = \begin{cases} \sqrt{\frac{1}{N}}, & k = 0 \\ \sqrt{\frac{2}{N}}, & k = 1,2,...N-1 \end{cases}$$

Inverse Discrete Cosine Transformation is defined as equation(5).

$$x(m,n) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} a(u)a(v)y(u,v)\cos\left[\frac{(2m+1)u\pi}{2M}\right]\cos\left[\frac{(2n+1)v\pi}{2N}\right] \tag{5}$$

In general, DCT dimensionality reduction choose those big coefficients at top left corner as feature. And there exists many DCT coefficients selection method. Figure 3 showes three popular method: square, triangular and circular selection method.

But the above three selection methods are all by hand, can these DCT coefficient be selected automatically ? In geometry, all samples form a elliptical sphere of N dimensionality. the eigenvectors of covariance matrix are

axis of this elliptical sphere cloud. PCA realize dimensionality reduction by finding these dominating axis. Enlightened by this idea,if we build a covariance matrix based on DCT coefficients,PCA can be used to extract the DCT coefficients automatically.

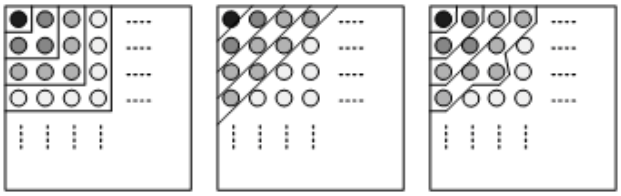

Figure 3. Three DCT coefficients selection method. From left to right they are square,triangular and circular selection method respectively.

III. THEORY OF LDAO ALGORITHM

LDA maximizes the ratio of between-class variance to the the within-class variance in any particular data set thereby guaranteeing maximal separability. The prime difference between LDA and PCA is that PCA does more of feature classification and LDA does data classification. In PCA, the shape and location of the original data sets changes when transformed to a different space whereas LDA doesn't change the location but only tries to provide more class separability and draw a decision region between the given classes.This method also helps to better understand the distribution of the feature data.

LDA is based on Fisher criterion function, see equation (6), it aims at searching the most discriminant projection direction matrix W for feature dimension reduction. For multiple classification, S_b 、 S_w stands for between-class scatter matrix and within-class scatter matrix respectively. LDA maximizes the ratio of between-class variance to the within-class variance thereby guaranteeing maximal separability.

$$J(W) = \arg\max_{W} \frac{|W^T S_b W|}{|W^T S_w W|} \tag{6}$$

Figure 4 is the class selection direction comparison between traditional LDA and LDAO.Speech recognition or speechreading is actually matter of classification, Let's assume C objects is to be classified. In LDA if we follow traditional class unit selection direction, that is, phoneme, syllable, sub-syllable or HMM state would be used as class unit, then the produced projection matrix W only ensures feature dimension reduction follow the direction which is most discriminant among these class units. Unfortunately, this projection direction doesn't have direct relations with recognition accuracy. So the traditional LDA algorithm is not the optimal dimension reduction strategy.

Secondly, according to Bayes maximum posterior probability theory, the optimal classification criterion is based on equation (7):

$$k^* = \arg\max_{1\le k\le C} P(M_k \mid X) = \arg\max_{1\le k\le C} \frac{P(X \mid M_k)P(M_k)}{P(X)} \tag{7}$$

where $P(X)$ is a constant, $P(M_k)$ is the apriori probabilities of the classes. To simply the theory discussion, the apriori probabilities is assumed to be equal. So equation (7) can be expressed as follows:

$$k^* = \arg\max_{1\le k\le C} P(X \mid M_k) \tag{8}$$

C is the number of total objects. If HMM is used as recognition model, M_k is the HMM of the *k*th object, X is a sequence of feature vectors, Because Baum training algorithm of HMM is sub-optimal, that is, it only maximize the output probability of the reduced feature vector on this trained HMM, but can't guarantee the best classification among objects.

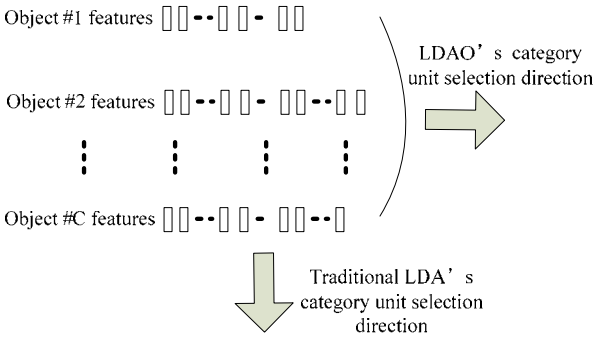

Figure 4. The class unit selection direction comparison between traditional LDA and LDAO

In contrast, LDAO exploit all objects to be classified as class unit, S_b 、 S_w are trained on these data belonging to different class units, so the obtained projection matrix W denotes the best direction of dimension reduction. That is, dimension reduction follows the direction of optimal recognition accuracy. Feature extraction process based on LDAO is as follows:

Assuming C objects (Chinese word) are to be classified, ROI size is 24×24, DCT is performed on these ROI in advance, and retain the bigger 55 DCT coefficients at top left corner of ROI. Then every word is normalized to 30 frames of the same length. Because the feature dimension of each word X is d=55×30, which is much bigger than the number of total samples, result in matrix S_w singular, that is:

$$Rank(S_W) \le N - C \tag{9}$$

N denotes the number of total samples. So the feature dimension X must be reduced before LDA, in general, PCA is fit for this task: first, build covariance matrix A based on all samples, and KL transformation is performed on covariance matrix A, then the obtained matrix $W_{pca,}$ is used to perform linear transformation to feature vector X based on equation (10), feature dimension is reduced to

$d1(d1 \le N-C)$ from d, forming a new feature Z in subspace.

$$Z_k = W_{PCA}^T X_k \qquad (k = 1,2,...,N) \tag{10}$$

For low dimension feature Z, LDA can be performed directly. Now redefine the between-class scatter matrix S_b and within-class scatter matrix S_w as equation (11), (12):

$$S_w = \frac{1}{N}\sum_{i=1}^{C}\sum_{k=1}^{N_i}(x_{ik}-\mu_i)(x_{ik}-\mu_i)^T \tag{11}$$

$$S_b = \frac{1}{N}\sum_{i=1}^{C}(\mu_i-\mu)(\mu_i-\mu)^T \tag{12}$$

where N_i stands for the number of the i th class samples, μ_i is the mean of the i th class samples. μ is the mean of total samples, :

$$\mu_i = \frac{1}{N_i}\sum_{k=1}^{N_i} x_{ik} \tag{13}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{C} N_i\mu_i \tag{14}$$

feature extraction based on LDA is actually projecting feature from space R^{d1} to low dimension space R^{C-1}:

$$Y_k = W_{LDA}^T Z_k \qquad (k = 1,2,...,N) \tag{15}$$

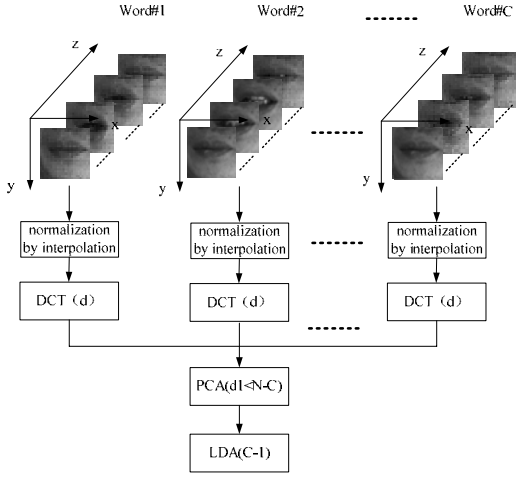

Figure 5. Feature extracting process based on LDAO, what in bracket is feature dimension

Finally, projection matrix W_{LDA} of LDA can be obtained when maximizing the criterion function $J(W)$ in equation(6). $J(W)$ is a generalized Rayleigh entropy. Matrix W_{LDA} can be obtained by the use of Lagrange multiplier method. As matrix S_w is nonsingular, so W_{LDA} can be solved by equation (16) . that is, calculate eigenvalue of matrix $S_w^{-1}S_b$, then rank these eigenvalue according to value. select M eigenvectors corresponding to former M bigger eigenvalue and build projection matrix W_{LDA}.

$$S_W^{-1}S_bW_{LDA} = \lambda W_{LDA} \tag{16}$$

Based on the obtained two linear transformation matrix W_{PCA} and W_{LDA}, feature extraction in speechreading can be expressed in one equation (17) . The integrated feature extraction process based on LDAO is shown in Figure 3.

$$Y_k = W_{PCA}^T W_{LDA}^T X_k \qquad (k = 1,2,...,N) \tag{17}$$

IV. TRAINING METHOD OF LDAO

As all isolated words have been normalized to the same length in advance, that is, each word has the same feature dimension. so there's no need to use complex HMM which is suitable for handling time-varying sequence. Euclidean distance can be used to classify data now. but Euclidean distance neglects the variance of different class. Here we use Gauss Mixture Model (GMM). Actually, GMM is a kind of multi-dimension probability density function, a GMM with M mixture unit can be expressed by equation (18):

$$P(x) = \sum_{i=1}^{N} w_i \frac{\exp[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)]}{(2\pi)^{1/2}|\Sigma_i|^{1/2}} \tag{18}$$

μ_i、$\sum_i$ stand for the mean and covariance of the i th mixture unit in GMM respectively, N is the number of mixture unit, w_i is the weight of ith mixture unit, so the model of one word can be expressed as equation (19):

$$\lambda = (w_i, \mu_i, \Sigma_i) \qquad (i = 1,2,...,M) \tag{19}$$

As each mixture unit can stands for features of different people. So GMM can realize speaker-independent speechreading compared with Euclidean distance method. Expectation Maximization (EM) algorithm can be used to train this model[26]:

Assuming feature vector $x_t(1,2,...,T)$ are used to train GMM. Maximum Likelihood Estimate is usually introduced to estimate this unknown parameters of known distribution function. the trained GMM parameters should maximize the mean output probability of $x_t(1,2,...,T)$, that is, maximizing the following Log-likelihood function:

$$L = P(X \mid \lambda) = \sum_{t=1}^{T}\log P[x_t \mid \lambda] \tag{20}$$

Assuming original GMM is λ, training goal is estimating a new GMM $\overline{\lambda}$ based on equation (21):

$$P(X \mid \overline{\lambda}) \geq P(X \mid \lambda) \tag{21}$$

An auxiliary function $Q(\lambda, \overline{\lambda})$ is introduced as follows:

$$Q(\lambda, \overline{\lambda}) = \sum_{I} p(X, I \mid \lambda) \log p(X, I \mid \overline{\lambda}) \tag{22}$$

I is one of the state sequences, new GMM parameters $\overline{\lambda}$ can be obtained through maximizing auxiliary function $Q(\lambda, \overline{\lambda})$, the iterative equation for the weight, mean and covariance of GMM are as follows:

$$\overline{w_i} = \frac{1}{T} \sum_{t=1}^{T} P(i \mid x_t, \mu_i, \Sigma_i) \tag{23}$$

$$\overline{\mu_i} = \frac{\sum_{t=1}^{T} P(i \mid x_t, \mu_i, \Sigma_i) x_t}{\sum_{t=1}^{T} P(i \mid x_t, \mu_i, \Sigma_i)} \tag{24}$$

$$\overline{\Sigma_i} = \frac{\sum_{t=1}^{T} P(i \mid x_t, \mu_i, \Sigma_i)(x_t - \mu_i)(x_t - \mu_i)^T}{\sum_{t=1}^{T} P(i \mid x_t, \mu_i, \Sigma_i)} \tag{25}$$

To recognize an unknown new speechreading samples, firstly extracting feature vector by equation (17), then input this low dimension feature into each GMM to compute the output probability, the GMM corresponding to the biggest output probability value is the recognized result.

V. EXPERIMENTAL RESULTS AND ANALYSIS

Three experiments were performed: the first is making a comparison between several traditional appearance-based method. The second is to test classification performance of LDAO, a comparison between LDAO and DCT+LDA which is recognized as currently the best algorithm was performed; the last is to discover the relation between the number of reduced feature dimension and recognition accuracy. Experiments were performed on Matlab 7.0 + VC 6.0 platform, speechreading data was collected on VC 6.0, and all kinds of algorithm were executed on Matlab 7.0.

Three experiments were performed: the first is making a comparison between several traditional appearance-based method. The second is to test classification performance of LDAO, a comparison between LDAO and DCT+LDA which is recognized as currently the best algorithm was performed; the last is to discover the relation between the number of reduced feature dimension and recognition accuracy. Experiments were performed on Matlab 7.0 + VC 6.0 platform, speechreading data was collected on VC 6.0, and all kinds of algorithm were executed on Matlab 7.0.

A. Database

To access the quality of LADO algorithm, we build a bimodal speechreading database, which contains 30 person. The recognition task is 10 Chinese command words and 0~9 ten digital words, each word is repeated 12 times. The character of this database is video camera shooting ROI directly, so ROI location error is avoided. The ROI is 24 bit color picture,and size is fixed to 160× 160. using cross trainning strategy, that is, 6 samples are used for trainning and the other 6 samples for recognition for each word. With the help of premiere software Speechreading data were obtained by Aligning audio channel to the video frames. Every ROI is downsampled to size 24 × 24 and grayed before extracting feature.

B. Comparison between PCA and DCT

1. Feature normalization

To eliminate the illumination effect,feature normalization must be done. Potamianos proposed a kind of normalization method before PCA[25]:

$$R = \frac{1}{N} \sum_{i=1}^{N} \frac{(g_{i,k} - m_k)}{\sigma_k} \frac{(g_{i,k} - m_k')}{\sigma_k'} \tag{26}$$

Where N is the number of samples, m_k、σ_k is the mean, variance of i th dimension of the feature. But experiments showed this method couldn't improve recognition accuracy distinctly, this method only normalize feature to (-3,3).

in our experiments we adopted two normalization steps.the first is histogram equalization, which is fit for eliminating contrast difference under varying illumination; the second step is normalizing feature value in transformed domain.for example, In PCA subspace, assuming the k th dimensionality feature of ROI is g_k ,then normalizing g_k to (0,1) based on equation (27)

$$g_k = \frac{g_k - \min(g_k)}{\max(g_k) - \min(g_k)} \tag{27}$$

To visualize normalization effect, In figure 6 we use the biggest eigenvalue of PCA as one dimension coordination. it shows the comparison between the normalized feature and original feature. In figure 7 we use the biggest two eigenvalue as coordination.

Experiments show that histogram equalization improve recognition accuracy about 15% , feature value normalization improve recognition accuracy about 10%.

2. Dynamic Differential Feature

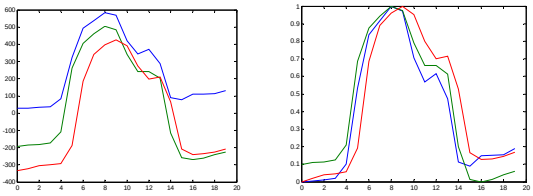

Figure 6. Comparison between original feature(left) and normalized feature(right) of Chinese word “qian” repeated 3 times at different time.

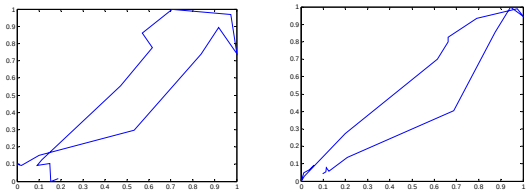

Figure 7. The effect of normalization on two Chinese word “qian” spoken at different time in 2-D feature domain

But obtained g_k is static feature. In fact, speechreading is based on the movement of lip, dynamic differential feature should be a more appropriate representation for speechreading feature. The calculation equation of first order difference feature is as follows:

$$d(n)=\frac{1}{\sqrt{\sum_{i=-k}^{k} i^2}}\sum_{-k}^{k} i\times c(n+i) \qquad (28)$$

Where c and d denote feature parameters, k stands for the range of differential feature, k is 2 in our experiments. at last, connect static feature with dynamic feature and create a new feature vector for recognition.

In recognition experiments. HMM is introduced as classifier, Which contain 6 states, Each state is a GMM with two mixture gauss function. From Table 1 to Table 3 are the experimental results of PCA, DCT, DCT+PCA respectively.

TABLE I. RECOGNITION ACCURACY BASED ON PCA

PCA	V	2	6	18	30	54
	accuracy	22%	86%	83%	78%	78%
	V+△V	2+2	6+6	18+18	30+30	54+54
	accuracy	32%	89%	85%	83%	83%

TABLE II. RECOGNITION ACCURACY BASED ON DCT

DCT	V	9	27	54	63	72
	accuracy	82%	90%	91%	89%	88%
	V+△V	9+9	27+27	54+54	63+63	72+72
	accuracy	87%	93%	93%	92%	92%

TABLE III. RECOGNITION ACCURACY BASED ON DCT+PCA

DCT+PCA	V+△V	2+2	6+6	18+18	30+30	54+54
	accuracy	85%	90%	89%	89%	88%

Experiments showed that dynamic differential feature improve recognition accuracy clearly. DCT achieved the best recognition results among three methods. But DCT+PCA doesn’t improve DCT accuracy. We think this is due to the rationale of HMM classifier. As the transformation matrix A of PCA is produced on all samples, That is, MMSE is relative to all samples. but this will lose information when A is performed on a ROI of a word, on the contrary, the transformation matrix of DCT is just derived from every ROI itself. So DCT retain feature of each ROI more perfectly. HMM is a kind of Statistical model, so losing too much original information when dimension reduction will influence the recognition accuracy of HMM.

C. Recognition Accuracy Comparison Between LDAO and Traditional LDA

As the length of each word is different ,but LDAO algorithm can only handle the fixed dimension feature vector, So each word must be normalized to the same length first. In general, speaking one Chinese word takes less than 1.2 seconds at normal speaking rate. So speaking frame length of any word would be less than 30 when frame rate is 25, In experiments all words were resampled to the length of 30 frames based on Spline Interpolation in advance.

In the first experiment, two kinds of traditional LDA algorithm which use semi-syllable and HMM state as class unit were introduced to compare with LDAO, see figure 8. Training and recognition were all based on HMM, which comprise 6 states, each state is a GMM with two mixture gauss function. In contrast, LDAO is based on GMM which has 2 mixture gauss function.

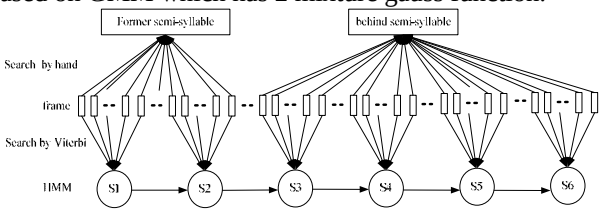

Figure 8. Two traditional feature extraction method based on LDA, one is using semi-syllable as class unit, the other is using HMM states as class unit.

When semi-syllable is used as the class unit in LDA, each word is divided into two parts according to 1:2 proportion, the former part denotes consonant, The second part stands for vowel which last longer time than consonant in speaking. every frame vector is assigned to two parts accordingly. Btween-class scatter matrix S_b and within-class scatter matrix S_w are trained on these data. And projection direction matrix W can be obtained by equation (6). When HMM state is used as the class unit in LDA, all frame vectors are assigned to six HMM states averagely, Baum algorithm is used to train HMM, then based on Viterbi algorithm find the optimal frame vector corresponding to each HMM state. Then the optimal projection direction matrix W can be found on these data.

To overcome the unreasonable assumption that every feature vector is irrelated in HMM theory. We made super-frames with a sequence of adjacent frame vectors. That is, connecting current frame with former t frames and latter t frames and forming a super-frame feature vector.

Table IV shows that when using HMM state as class unit the best recognition accuracy is 93.8%. which is slightly better than 93.6% when using semi-syllable as class unit. It's proved that Viterbi algorithm played a key role in assigning frame vectors to different semi-syllables. Viterbi algorithm is more accuracy in segmenting frames to semi-syllable than by hand. On the other hand, when increasing the number of super-frames, both two method improved the recognition accuracy. furthermore, when super-frame number is increased to 3, recognition accuracy enhanced obviously, and when super-frame is further increased , recognition accuracy just improved slightly. It showes that neighbouring frames are correlative, and this correlative property fades gradually as the distance of the frames increase. Compared with above two traditional LDA algorithm, LDAO algorithm showed great superiority. The best recognition accuracy of LDAO is 96.7%. Which is better than traditional DCT+LDA about 3%.

Although Potamianos performed the third linear transformation MLLR after DCT+LDA in his feature extraction method, see Figure 1. but MLLR only perform auto-adaptive for speaker-independent recognition task. In our speaker-dependent speechreading recognition experiments MLLR can be omitted. This won't influence the comparison results.

TABLE IV. SPEECHREADING RECOGNITION ACCURACY COMPARISON BETWEEN LDAO AND TRADITIONAL DCT+LDA

Algorithm	Number of super-frames			
	1	3	5	7
DCT+LDA based on semi-syllable	93.2%	93.5%	93.6%	93.6%
DCT+LDA based on HMM state	93.6%	93.8%	93.8%	93.8%
LDAO	96.7%			

D. Dimension reduction performance of LDAO

The lower feature dimension is, the better a algorithm's real-time property would be. In the second experiment, the dimension reducing performance of LDAO is tested. because choosing HMM state as class unit is better than other class units in traditional LDA, so HMM state is selected as class unit in the following traditional LDA experiments.

Figure 9 gives the dimension reducing capability comparison between DCT, PCA, DCT+LDA and LDAO. When feature extraction is based on PCA, the best recognition accuracy is 89% with 6 static feature vector and 6 dynamic feature vector; For DCT, the best recognition accuracy is 93% with 27 static feature vector and 27 dynamic feature vector, apparently, DCT is better than PCA; When based on DCT+LDA with 3 super-frames, the best recognition accuracy is 93.8% with 9 static feature vector plus 9 dynamic feature vector; by comparison, LDAO achieve the best result, the recognition accuracy is 96.7% with only 8 feature vector. Obviously, although DCT+LDA achieve the best results among traditional feature extraction algorithms, LDAO is better than DCT+LDA about 3%.

It must be stated that the above mentioned feature dimensionality of DCT, PCA and DCT+LDA only belong to single frame of a word, on the contrary, feature dimension of LDAO belongs to a whole word.

Obviously, the extracted feature dimension based on LDAO is much more lower than traditional appearance-based algorithms. In Figure 9, The LDAO recognition accuracy reach 94.5% with only 4 dimension feature, and achieve the best accuracy 96.7% when feature dimension is 8. then the recognition accuracy is almost stable as feature dimension increase to 19. In conclusion, LDAO not only showed the excellent classification property, but also exhibited the best feature dimension reducing capability.

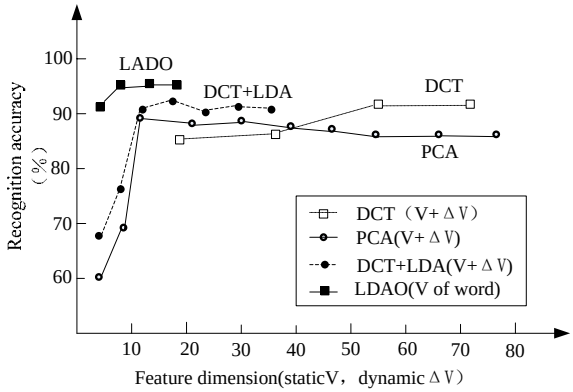

Figure 9. The reducing dimension capability comparison between DCT, PCA, DCT+LDA and LDAO

Furthermore, concerning the model complexity of training and recognizing , DCT,PCA and DCT+LDA are all based on HMM. LDAO is based on GMM, which is much more simple than HMM in training. So LDAO algorithm is more fit for real-time application.

VI. *CONCLUSION*

To solve the problem of extracting feature in speechreading, A new LDAO algorithm which is fit for isolated word recognition is introduced in this paper. Unlike traditional LDA algorithm which use phoneme, syllable,HMM states as class unit, LDAO use the objects to be recognized as class unit to linear discriminant analysis. So in theory LDAO algorithm guarantees feature extracting follows the optimal recognition direction. And overcomes the shortcoming that feature extracting direction is irrelative to recognition accuracy in traditional LDA. Furthermore,To meet with speaker-independent speechreading recognition task, GMM was introduced to act as training and recognition model.To test LDAO algorithm, DCT, PCA and DCT+LDA were compared on speaker-dependent bimodal database. Experimental results showed that LDAO achieved the best recognition accuracy with the least dimensionality among all appearance-based algorithms. Of course, as LDAO use object as class unit, LDAO algorithm is only fit for recognizing limited isolated word.

ACKNOWLEDGMENT

This work was supported in part by a grant from education fund of Jiangxi(No.GJJ08012) and innovation fund of NanChang University.

REFERENCES

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent Advances in the Automatic Recognition of Audiovisual Speech," Proc. IEEE Int'l Conf. Image Processing, vol. 91, no. 9, pp. 1306-1326, Sept. 2003.

[2] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in Proc. Workshop Multimedia Signal Processing, 1998, pp. 65–70.

[3] A.Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, 1996, pp. 461–471.

[4] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in Proc. Workshop Multimedia Signal Processing, 1998, pp. 65–70.

[5] Y.Zhang,S. Levinson, and T. Huang, "Speaker independent audiovisual speech recognition," in Proc. Int. Conf. Multimedia Expo,2000, pp. 1073–1076.

[6] A.J.Goldschen, O. N. Garcia, and E. D. Petajan, "Rationale for phoneme-viseme mapping and feature selection in visual speech recognition," in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag, 1996, pp. 505–515.

[7] Wang, S.L. Liew, A.W.C. "An automatic lipreading system for spoken digits with limited training data" IEEE Transactions on Circuits and Systems for Video Technology, v 18, n 12, p 1760-1765, December 2008.

[8] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition," in Proc. Int. Conf. Acoustics, Speech, and Signal Processing, 2001, pp. 177–180.

[9] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour model," Int'l J. Computer Vision, vol. 1, pp. 321-331, 1988..

[10] He, Jun. ZhangHua. "Research on visual automatic speech recognition". Journal of Information and Computational Science, v 5, n 4, p 1567-1575, July 2008..

[11] J. R. Movellan and G. Chadderdon, "Channel separability in the audio visual integration of speech: A Bayesian approach," in Speechreading by Humans and Machines, D. G. Stork and M. E.Hennecke, Eds. Berlin, Germany: Springer-Verlag, 1996, pp.473–487.

[12] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE Trans. Multimedia, vol. 2, pp.141–151, Sept. 2000.

[13] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," EURASIP J. Appl. Signal Process., vol. 2002 , pp. 1274–1288, Nov. 2002..

[14] Bregler andY.Konig, "'Eigenlips' for robust speech recognition," in Proc. Int. Conf. Acoustics, Speech, and Signal Processing, 1994,pp. 669–672.

[15] Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri,J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition,"Center Lang. Speech Process., Johns Hopkins Univ., Baltimore,MD, 2000.

[16] P. Scanlon and R. Reilly, "Feature analysis for automatic speechreading," in Proc. Workshop Multimedia Signal Processing,2001, pp. 625–630.

[17] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp.141–151, Sept. 2000.

[18] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 2002 , pp. 1274–1288,Nov. 2002.

[19] Matthews,I.Cootes,T.F.Bangham,J.A.Cox,S.Harvey,R. "Extraction of visual features for lipreading". Pattern Analysis and Machine Intelligence, IEEE Transactions. Feb 2002. Volume: 24, Issue: 2.p 198-213.

[20] M. T. Chan, "HMM based audio-visual speech recognition integrating geometric- and appearance-based visual features," in Proc.Workshop Multimedia Signal Processing, 2001, pp. 9–14.

[21] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in Proc. Eur. Conf. Computer Vision, 1998, pp. 484–498.

[22] C. Neti, G.Potamianos, J. Luettin, I. Matthews, H.Glotin, D.Vergyri. "Large-vocabulary audio-visual speech recognition: a summary of theJohns Hopkins Summer 2000 Workshop". Multimedia Signal Processing, 2001 IEEE Fourth Workshop on. 2001p : 619-624.

[23] R Haeb-Umbach, H Ney, Linear Discriminant Analysis for improved large vocabulary continuous speech recognition[C].In:Proceedings of ICASSP 92,1992. 1: 13 ~ 16.

[24] Sakai, Makoto. Kitaoka, Norihide; Nakagawa, Seiichi. "Generalization of linear discriminant analysis used in segmental unit input HMM for speech recognition". ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, v 4, p IV333-IV336, 2007.

[25] Potamianos, G.; Graf, H.P. Linear discriminant analysis for speechreading.Multimedia Signal Processing, 1998 IEEE Second Workshop on Volume , Issue , 7-9 Dec 1998 Page(s):221 – 226.

[26] Laird, N.M., Lange, N., and Stram, D. "Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm," Journal of the American Statistical Association, 82, 97 -105.

Jun He, born in 1969. Received his B.A's and M.A's degrees in Control Theory and Control Engineering in NanChang University, in 1991 and 1999 respectively. Since 2005, he has been a vice-professor and PH.D. degree candidate in Electronic and Mechanic engineering in NanChang University. His current research interests inlude pattern recognition, speech recognition, human-machine interaction.

Hua Zhang , born in 1964, Professor and Ph. D. supervisor in NanChang University, director of Chinese Welding Association, vice chairman of Chinese Robot and Automation Committee. Member of IEEE. He received his Ph. D. degree in automatic welding in Tsinghua University. more than 25 papers were published in《Science in China》,《Welding Journal》. his Research area: automatic welding and moving robotic welding system, intelligent service robot.