

LINEAR DISCRIMINANT ANALYSIS FOR SPEECHREADING

Gerasimos Potamianos and Hans Peter Graf
AT&T Labs-Research, Florham Park, NJ 07932-0971
{makis,hpg}@research.att.com

Abstract - This paper investigates the use of Fisher-Rao linear discriminant analysis (LDA) as a means of visual feature extraction for hidden Markov model based automatic speechreading. For every video frame, a three-dimensional region of interest containing the speaker's mouth over a sequence of adjacent frames is lexicographically arranged into a data vector. Such vectors are then projected onto the space of the most discriminant "eigensequences", estimated by means of LDA on a training set of image sequence vectors, labeled from a set of a-priori chosen classes. The resulting projections, as well as their first and second derivatives over time, are used as features for automatic speechreading. The proposed method is applied to single-speaker, multi-speaker, and speaker-independent visual-only recognition tasks, consistently outperforming principal component analysis and discrete wavelet transform based visual features. Specific issues relevant to LDA are also discussed, namely, class selection, automatic data class labeling, and dimensionality reduction prior to LDA.

INTRODUCTION

One of the main difficulties in automatic *speechreading* is the *visual front end* design. Given the image sequence of a speaker's mouth region, such a design should extract visual features that contain relevant information about the spoken word sequence and allow high accuracy, speaker-independent speechreading by means of a *hidden Markov model* (HMM) based recognizer [1]. Such features can then be used in addition to traditional audio features to improve automatic speech recognition [2], [3].

In general, speechreading features are based on lip contour modeling [3], or on compressed representations of the image pixel values containing the speaker's mouth [2-5]. Various such representations are studied in [5]. Among them, the *discrete wavelet transform* (DWT) [5], as well as a *principal component analysis* (PCA) based projection [2], [6], yield high accuracy visual-only recognition, consistently outperforming lip contour based features. However, neither the DWT, nor PCA, is designed to achieve optimal discrimination among the classes of interest, such as the spoken words. Fisher-Rao *linear discriminant analysis* (LDA) [7] provides the optimal data projection for achieving maximum such discrimination. Not surprisingly, LDA has been

successful in a variety of classification tasks, such as image retrieval [8], face recognition [9], visual-only speaker identification [10], and traditional automatic speech recognition [11]. In the context of speechreading, LDA has to date only been considered for a single-speaker task [4].

In this paper, we apply LDA to both single-speaker and speaker-independent speechreading, and we investigate factors that affect performance of the resulting speechreading system. We propose the use of LDA on data corresponding to a three-dimensional *region of interest* (ROI) that contains the speaker's mouth over a sequence of adjacent frames, and we experimentally demonstrate that the resulting visual features yield significantly higher visual-only recognition accuracy than both DWT and PCA based features. Performance is maximized when all HMM states are considered as separate classes. Finally, we investigate the use of a PCA projection prior to the use of LDA, as a means of dimensionality reduction [8], [10]. We conclude that such a projection is undesirable in our case.

DATA PROJECTIONS FOR FEATURE EXTRACTION

Let $\{g(x, y, t) : 1 \leq x \leq M, 1 \leq y \leq N, 1 \leq t \leq T\}$ denote an image sequence of the speaker's mouth region, preprocessed as in [5]. At every time instance t , we consider as data relevant to speechreading the MNK pixels in $\{g(x, y, z) : 1 \leq x \leq M, 1 \leq y \leq N, t - t_L \leq z \leq t + t_R\}$, where $t_L, t_R \geq 0$, and $K = t_L + t_R + 1$. We lexicographically order this data into an MNK -element data vector $\underline{g}_t = [g_{t,1}, \dots, g_{t,MNK}]'$. We then seek a *projection* matrix $\mathbf{P} = [\underline{P}_1, \dots, \underline{P}_J]$ of size $MNK \times J$, where $J \ll MNK$, such that the J -element feature vector $\underline{o}_t = \mathbf{P}' \underline{g}_t$ contains most speechreading information. To obtain \mathbf{P} , we are given L training examples $\{\underline{g}_l, l = 1, \dots, L\}$, where typically $L \gg MNK$.¹

Linear discriminant analysis LDA [7] assumes that a set of classes $\mathcal{C} = \{1, 2, \dots, |\mathcal{C}|\}$ is a-priori given, as well as that the training set data vectors \underline{g}_l , $l = 1, \dots, L$, are labeled as $c(l) \in \mathcal{C}$. LDA seeks a projection \mathbf{P}_{LDA} , such that the projected training sample $\{\mathbf{P}'_{\text{LDA}} \underline{g}_l, l = 1, \dots, L\}$ is "well separated" into the set of classes \mathcal{C} . Formally, \mathbf{P}_{LDA} maximizes

$$Q(\mathbf{P}) = \frac{\det(\mathbf{P}' \mathbf{S}_B \mathbf{P})}{\det(\mathbf{P}' \mathbf{S}_W \mathbf{P})}, \quad (1)$$

where $\det(\bullet)$ denotes matrix determinant, and \mathbf{S}_W , \mathbf{S}_B are the *within-class scatter* and *between-class scatter* matrices of the training sample, defined as

$$\mathbf{S}_W = \sum_{i=1}^{|\mathcal{C}|} P(i) \Sigma^{(i)} \quad \text{and} \quad \mathbf{S}_B = \sum_{i=1}^{|\mathcal{C}|} P(i) (\underline{\mu}^{(i)} - \underline{\mu})(\underline{\mu}^{(i)} - \underline{\mu})', \quad (2)$$

respectively. In (2), $P(i) = L_i/L$, $i = 1, \dots, |\mathcal{C}|$, denotes the class empirical probability mass function, where $L_i = \sum_{l=1}^L \delta_{c(l)}^i$, and $\delta_i^j = 1$, if $i = j$; 0,

¹ In our case, $M = N = 16$, $1 \leq K \leq 5$, and $L = O(10^5)$.

otherwise. In addition, for each class $i=1, \dots, |\mathcal{C}|$, its sample mean is

$$\underline{\mu}^{(i)} = [\mu_1^{(i)}, \dots, \mu_{MKN}^{(i)}]', \quad \text{where } \mu_k^{(i)} = \frac{1}{L_i} \sum_{l=1}^L \delta_{c(l)}^i g_{l,k}, \quad (3a)$$

and the class sample *covariance* is $\Sigma^{(i)}$, with elements given by

$$\sigma_{k,k'}^{(i)} = \frac{1}{L_i} \sum_{l=1}^L \delta_{c(l)}^i (g_{l,k} - \mu_k^{(i)}) (g_{l,k'} - \mu_{k'}^{(i)}), \quad (3b)$$

for $k, k' = 1, \dots, MKN$. Finally, $\underline{\mu} = \sum_{i=1}^{|\mathcal{C}|} P(i) \underline{\mu}^{(i)}$ is the total sample *mean*.

To maximize (1), the *generalized eigenvalues* and right *eigenvectors* of the matrix pair $(\mathbf{S}_B, \mathbf{S}_W)$ are computed that satisfy $\mathbf{S}_B \mathbf{A} = \mathbf{S}_W \mathbf{A} \boldsymbol{\Lambda}$ [7], [12]. Let the J largest generalized eigenvalues be $\{\lambda_{k_1}, \dots, \lambda_{k_J}\}$, where $\lambda_{k_j} = \mathbf{A}_{k_j k_j}$, and let matrix $\mathbf{A} = [\underline{a}_1, \underline{a}_2, \dots, \underline{a}_{MKN}]$ have as columns the generalized eigenvectors. Then, $\mathbf{P}_{LDA} = [\underline{a}_{k_1}, \dots, \underline{a}_{k_J}]$. Given data \underline{g}_t , we extract its J -element feature vector

$$\underline{o}_t = [o_{t,1}, \dots, o_{t,J}]', \quad \text{as } o_{t,j} = \langle \underline{g}_t, \underline{a}_{k_j} \rangle, \quad (4)$$

where $\langle \bullet, \bullet \rangle$ denotes vector inner product.

Vectors \underline{a}_{k_j} , for $j = 1, \dots, J$, are the linear discriminant “eigensequences” that correspond to the directions where the data vector projection yields high discrimination among the classes of interest. In our experiments, we have considered various class choices \mathcal{C} , as we discuss in our Results Section. We should note that the rank of \mathbf{S}_B is at most $|\mathcal{C}| - 1$, hence we consider $J \leq |\mathcal{C}| - 1$, whereas the rank of \mathbf{S}_W cannot exceed $L - |\mathcal{C}|$. Thus, in case of insufficient training data, a PCA projection is typically used before applying LDA, in order to achieve dimensionality reduction [8], [10].

Principal component analysis PCA [6] achieves optimal information compression, in the sense of minimum mean square error between \underline{g}_t and its reconstruction based on \underline{o}_t , by projecting the data onto the directions of their greatest variance. However, the problem of *scaling* arises when applying PCA to classification [6]. In our experiments, we found it beneficial to scale our data according to their inverse variance, and subtract their mean. Namely, we compute the data mean $m_k = E[g_{l,k}]$ and variance $\sigma_k^2 = E[g_{l,k} - m_k]^2$, for $k = 1, \dots, MKN$, and the data *correlation* $MNK \times MNK$ matrix \mathbf{R} with elements $E[(g_{l,k} - m_k)(g_{l,k'} - m_{k'})]/(\sigma_k \sigma_{k'})$, for $k, k' = 1, \dots, MKN$. We then *diagonalize* the correlation matrix, as $\mathbf{R} = \mathbf{B} \mathbf{D} \mathbf{B}'$ [12], where $\mathbf{B} = [\underline{b}_1, \underline{b}_2, \dots, \underline{b}_{MKN}]$ has as columns the eigenvectors of \mathbf{R} , and \mathbf{D} is the diagonal matrix containing the eigenvalues of \mathbf{R} . Let the J largest such eigenvalues be $\{\delta_{k_1}, \dots, \delta_{k_J}\}$, where $\delta_{k_j} = \mathbf{D}_{k_j k_j}$. Then, $\mathbf{P}_{PCA} = [\underline{b}_{k_1}, \dots, \underline{b}_{k_J}]$. Given a data vector \underline{g}_t , we first normalize it into

$$\underline{G}_t = [G_{t,1}, \dots, G_{t,MKN}]', \quad \text{where } G_{t,k} = \frac{g_{t,k} - m_k}{\sigma_k}, \quad (5a)$$

and then extract its J -element feature vector as

$$\underline{o}_t = [o_{t,1}, \dots, o_{t,J}]', \quad \text{where } o_{t,j} = \langle \underline{G}_t, \underline{b}_{k_j} \rangle. \quad (5b)$$

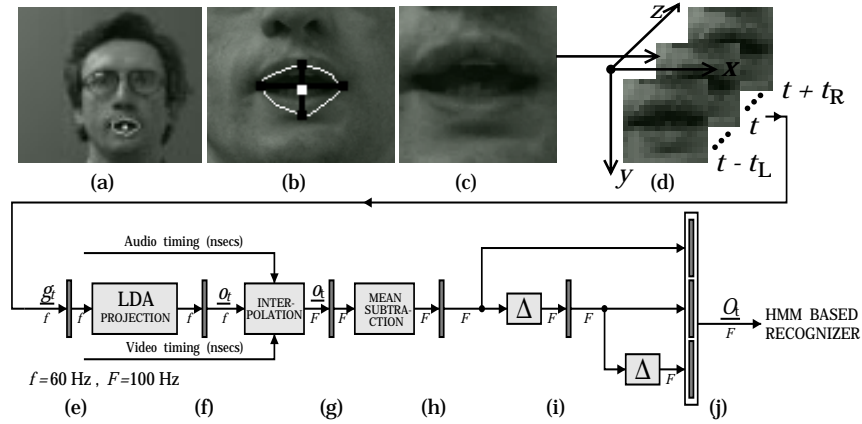


Figure 1: The visual front end. (a): Original video frame with estimated outer lip contour and mouth center [5]. (b): Speaker's mouth region, zoomed in. (c): Video frame ROI after an optional normalization step [5]. (d): Final, downsampled, three-dimensional ROI at time t . (e): Data vector \underline{g}_t . (f): LDA projection \underline{o}_t . (g): Feature interpolation to 100 Hz. (h): Feature mean subtraction. (i): Feature differentiation. (j): Final feature vector \underline{O}_t .

Often in practice, PCA precedes the use of LDA. In such a case, $\underline{o}_t = \mathbf{P}'_{\text{LDA}} \mathbf{P}'_{\text{PCA}} \underline{G}_t$, where \mathbf{P}'_{LDA} is computed by (2)-(4), but with the substitution $\underline{g}_t \leftarrow \mathbf{P}'_{\text{PCA}} \underline{G}_t$.

THE SPEECHREADING SYSTEM

The automatic speechreading system is described in detail in [5]. Briefly, at every video frame (available at a 60 Hz rate), data vector \underline{g}_t is first extracted, and then projected using (4) to create a "static" feature vector \underline{o}_t . Vector \underline{o}_t is postprocessed by a series of linear operations, namely, linear interpolation, mean subtraction, and first and second order differentiation, resulting to the final feature vector $\underline{O}_t = [\underline{o}_t - E[\underline{o}_t], \Delta \otimes \underline{o}_t, \Delta \otimes \Delta \otimes \underline{o}_t]$, which is now available at the audio feature rate of 100 Hz [5] (see Fig. 1).

As a means of statistical pattern matching, the speechreading system uses continuous density HMMs with observation probabilities modeled as multi-dimensional Gaussian mixtures with diagonal covariance matrices [1]. For the specific recognition tasks considered in this paper, context independent, whole word, 6-10 state, left-to-right HMMs with 16 mixtures per state, and a single state silence HMM with 32 mixtures are used. All HMM parameters are estimated by maximum likelihood Viterbi training [1]. The initial segmentation in this training procedure is obtained by using the audio channel information and maximum likelihood trained audio-only HMMs of *identical* topology to their visual counterparts.

In the case of LDA, the set of HMM states is first partitioned into classes. Then, aligning the audio channel forced segmentation [1] to the video frames provides a simple means of automatically labeling the image sequence training

Task	$ \mathcal{V} $	Train. set	Test set	DWT	PCA	LDA
P.1 - digits	11	$1 \times 400 \times 5$	$1 \times 200 \times 5$	92.4(67.5)	92.2(70.0)	95.7(83.0)
P.2 - letters	26	$1 \times 2000 \times 4$	$1 \times 500 \times 4$	57.4(11.8)	52.9 (9.0)	65.3(21.0)
P.3 - letters	26	$49 \times 20 \times 4$ $40 \times 25 \times 4$	$49 \times 5 \times 4$ $9 \times 25 \times 4$	29.6 (1.6) 19.6 (0.4)	31.3 (0.8) 18.9 (0.0)	36.5 (2.9) 20.0 (0.9)

Table 1: Speechreading experiments on three tasks of the AT&T bimodal database consisting of continuously spoken digit (P.1) and letter strings (P.2, P.3). A multi-speaker and a speaker-independent testing scenario are considered in task P.3, whereas both tasks, P.1 and P.2, are single-speaker. Training and test set sizes are given in number of speakers \times sentences per speaker \times words per sentence, and $|\mathcal{V}|$ denotes vocabulary size (Note: In task P.1, \mathcal{V} includes word “oh”). Unknown string length recognition results on the test set are depicted in word (string) % accuracies, using feature extraction by means of the DWT (see [5]) ($K=1, J=17$), PCA (see (5)) ($K=3, J=10$), and LDA (see (4)) ($K=5, J=18$).

vectors, and hence of computing matrices \mathbf{S}_W and \mathbf{S}_B , using (2) and (3).

RESULTS

Visual-only recognition experiments on three tasks of the AT&T audio-visual database are reported in this paper. In Table 1, the performance of LDA based features is compared with DWT [5] and PCA based features. Clearly, LDA consistently results in better speechreading. It is worth mentioning that the LDA results further improve by using discriminatively trained HMMs [1] to 97.0% (88.5%) and 68.2% (24.0%) word (string) accuracies for tasks P.1 and P.2, respectively.

In Fig. 2, we perform a number of LDA experiments on task P.1. Fig. 2(a) depicts speechreading performance as a function of the number J of DWT, PCA, and LDA “static” features used in the HMM automatic speechreading system. The superiority of LDA features is obvious. In Fig. 2(b), it becomes clear that the use of more than one successive video frames ($K > 1$) improves speechreading. In Fig. 2(c), the use of a PCA projection prior to LDA is investigated for the case $K = 2, J = 18$. Such a projection is clearly undesirable in our case. Finally, in Fig. 2(d), we depict the performance of LDA for various choices of classes.² We observe that clustering HMM states together into classes degrades speechreading performance.

SUMMARY AND DISCUSSION

In this paper, we use linear discriminant analysis on image sequences as a means of visual feature extraction for HMM based automatic speechreading. We demonstrate that LDA consistently achieves superior performance to both PCA and DWT based features on a variety of visual-only recognition tasks.

²In all cases, one class (denoted by sil) is dedicated to silence. In order to avoid its statistics dominating matrices \mathbf{S}_W and \mathbf{S}_B , we divide $P(\text{sil})$ by a large positive constant.

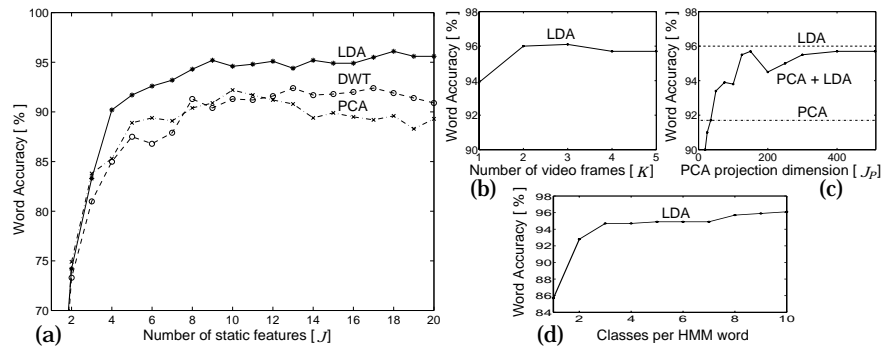


Figure 2: Speechreading recognition performance (word accuracy %) on database task P.1, as a function of: (a): Number J of “static” DWT ($K=1$), PCA ($K=3$), and LDA ($K=3$) based features. (b): Number K of successive frames in LDA ($J=18$). (c): Dimension of PCA projection based subspace prior to LDA application ($J=18$, $K=2$). (d): Number of classes per HMM word in LDA ($K=3$).

Automatic speechreading performance is best when the ROI consists of data from a small (2-5) number of successive frames as well as when all HMM states are considered as separate classes. Finally, when sufficient LDA training data exists, dimensionality reduction by means of a PCA data projection preceding LDA is undesirable.

References

- [1] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, 1993.
- [2] C. Bregler and Y. Konig, “Eigenlips’ for robust speech recognition”, *Proc. Int. Conf. Acoust. Speech Signal Process.*, Adelaide, pp. 669-672, 1994.
- [3] M.E. Hennecke, D.G. Stork, and K.V. Prasad, “Visionary speech: Looking ahead to practical speechreading systems”, in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 331-349, 1996.
- [4] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel, “Toward movement-invariant automatic lip-reading and speech recognition”, *Proc. Int. Conf. Acoust. Speech Signal Process.*, Detroit, pp. 109-112, 1995.
- [5] G. Potamianos, H.P. Graf, and E. Cosatto, “An image transform approach for HMM based automatic lipreading”, *Proc. Int. Conf. Image Process.*, Chicago, 1998.
- [6] C. Chatfield and A.J. Collins, *Introduction to Multivariate Analysis*. Chapman and Hall, London, 1980.
- [7] C.R. Rao, *Linear Statistical Inference and Its Applications*. John Wiley and Sons, New York, 1965.
- [8] D.L. Swets and J. Weng, “Using discriminant eigenfaces for image retrieval”, *IEEE Trans. Patt. Anal. Mach. Intel.*, Vol. 18, pp. 831-836, 1996.
- [9] K. Etemad and R. Chellappa, “Discriminant analysis for recognition of human face images”, in *Audio- and Video-based Biometric Person Authentication*, J. Bigün, G. Chollet, and G. Borgefors eds., Springer, Berlin, pp. 127-142, 1996.
- [10] T. Wark and S. Sridharan, “A syntactic approach to automatic lip feature extraction for speaker identification”, *Proc. Int. Conf. Acoust. Speech Signal Process.*, Seattle, pp. 3693-3696, 1998.
- [11] O. Siohan, “On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition”, *Proc. Int. Conf. Acoust. Speech Signal Process.*, Detroit, pp. 125-128, 1995.
- [12] G.H. Golub and C.F. Van Loan, *Matrix Computations*. The Johns Hopkins Univ. Press, Baltimore, 1983.