# MUTUAL INFORMATION EIGENLIPS FOR AUDIO-VISUAL SPEECH RECOGNITION

*Ivana Arsic and Jean-Philippe Thiran*

Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
email:{Ivana.Arsic, JP.Thiran}@epfl.ch
http:// itswww.epfl.ch

## ABSTRACT

This paper proposes an application of information theoretic approach for finding the most informative subset of eigenfeatures to be used for audio-visual speech recognition tasks. The state-of-the-art visual feature extraction methods in the area of speechreading rely on either pixel or geometric based methods or their combination. However, there is no common rule defining how these features have to be selected with respect to the chosen set of audio cues and how well they represent the classes of the uttered speech. Our main objective is to exploit the complementarity of audio and visual sources and select meaningful visual descriptors by the means of mutual information. We focus on the principal components projections of the mouth region images and apply the proposed method such that only those cues having the highest mutual information with word classes are retained. The algorithm is tested by performing various speech recognition experiments on a chosen audio-visual dataset. The obtained recognition rates are compared to those acquired using a conventional principal component analysis and promising results are shown.

## 1. INTRODUCTION

Research done in the area of audio-visual signal processing shows clear benefits of using a multi-modal approach for various tasks such as: speechreading, bimodal speaker recognition, speaker detection, etc. The overall system performance is improved when using the help of visual modality, especially in noisy and adverse environmental conditions. Still, the main issue is the choice of adequate visual features. Various types of visual cues as well as various methods for their extraction are proposed by the research community. A comprehensive overview can be found in [1].

Among those methods, area-based ones are of particular interest due to the stability and robustness. They are based on observing the whole mouth Region-of-Interest as a visual feature vector. Some of the widely used are Principal Component Analysis [2, 3] and the DCT transform [4]. No matter what kind of image transform is applied, selection of the visual features among the possible candidates is usually done following some a priori defined rule, inherited mainly from image compression. Thus, in case of DCT coefficients those with the highest energy are retained, while for eigenfeatures the choice is made upon the highest eigenvalues. Although the chosen features are suitable for image compression, it is not clear how "good" they are in terms of representing the class related information in multi-modal speech recognition. Since the only performance measure for speech recognition systems is recognition rate (or alternatively word error rate), a suitable measure should be used to show how optimal the chosen visual features are. A well known concept of mutual information can be employed for such tasks. The information theoretic approach for finding the most informative features for speech recognition purposes was previously used for audio-only feature selection in [5, 6]. Regarding the visual cues, to the best of our knowledge only the recent work by Scanlon et al. [7] considers the use of mutual information, as well as joint mutual information for feature selection tasks. The authors show the clear benefits of employing DCT coefficients chosen using mutual information for large-vocabulary visual speech recognition tasks.

In this paper we follow the same intuition that the most informative visual features would have high mutual information with respect to the speech classes. We consider the possibility of using the information theoretic framework and concept of mutual information for selecting the most informative principal components. The obtained mouth image projections on the maximum mutual information eigen space can be regarded as mutual information eigenlips. Further on we perform various visual-only and audio-visual isolated word recognition experiments, and compare the overall system performance to the one achieved using conventional principal component analysis.

The paper is organized as follows. In Section 2 we recall fundamentals of Principal Component Analysis (PCA) and describe the proposed information theoretic framework in detail. The experimental setup and obtained results are presented in Section 3, followed by conclusions and future work directions in Section 4.

## 2. METHOD

### 2.1 Principal component analysis

Principal Component Analysis (PCA) is a widely used data-driven approach for dimensionality reduction, optimal in the sense of information preservation. Regarding the area of visual speechreading this technique has been previously used to obtain a compact representation of mouth Region-of-Interest, known as *eigenlips* [3]. Visual features obtained employing this technique show to work well for the tasks of audio-visual speech recognition on different databases, [4, 8, 9].

The main idea behind a PCA approach is to project the data onto the directions of maximal variance. Thus, having

The work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2).

Figure 1: The first three principal components from the Tulips1 database.

a set of $N$ image training examples $\mathbf{x_1},\ldots,\mathbf{x_N}$, each image is considered as a $d$ dimensional feature vector ($d$ equals the number of pixels in an image).

First, an average image $\bar{\mathbf{x}}$ over the entire image set is found and subtracted from each image. Further on, $N$ image vectors are regarded as column vectors of the matrix $X$, having dimensions $d \times N$. Eigenvectors and eigenvalues are extracted from the covariance matrix of image data. The obtained eigenvectors are sorted on decreasing order of magnitudes of the corresponding eigenvalues. The most compact image set representation is obtained by keeping the number of principal components that account for $90-95\%$ of the variance. Figure 1 shows the first three eigenimages from the observed image set.

However, the retained eigenvectors capture the major variations across the training set (such as related to lighting directions), but carry no information about the relevance of selected features with respect to speech classes. Therefore, we propose to use an information theoretic approach and the basic principle of mutual information for selecting the relevant eigenfeatures.

### 2.2 Information Theoretic approach

Mutual information is a quantitative measure of the statistical dependence between two random variables [10]. The use of mutual information (MI) for finding the relevance of particular features with respect to a class is motivated by Fano's inequality [11] and data processing inequality. These well established concepts from information theory give a lower bound of the probability of error i.e. an upper bound of the probability of correct classification.

Let $y_i$ denote elements of visual feature vector as samples of the continuous random variable $Y$, such that $y_i \in R$, $i \in [1,N]$. Let $c_i$ denote class labels as samples of the discrete random variable $C$ such that $c_i \in \{1,\ldots,N_c\}$.

The mutual information $I(C;Y)$ is defined as:

$$I(C;Y) = H(C) - H(C|Y) = H(Y) - H(Y|C). \quad (1)$$

where $H(.)$ is Shannon's entropy in bits.

Since our focus is on audio-visual speech recognition, the classes of interest are those of the uttered speech. The entropy of a speech class $C$ is defined in terms of class prior probabilities $p(c)$ as:

$$H(C) = -\sum_c p(c) \cdot log_2 p(c). \quad (2)$$

Knowing the elements of the feature vector, the conditional entropy of a speech class given the feature vector can be expressed as:

$$H(C|Y) = -\int_y p(y) \cdot \left( \sum_c p(c|y) \cdot log_2 p(c|y) \right) dy. \quad (3)$$

Similarly, the conditional entropy of the feature vector given the class is:

$$H(Y|C) = -\sum_c p(c) \cdot \int_y p(y|c) log_2 p(y|c) dy. \quad (4)$$

The main problem when trying to apply mutual information criterion is the probability density function estimation. Given a practical dataset we can only find an approximation of the probability density function, and various parametric or non-parametric methods can be used for such purposes. Here, like in [7] a histogram-based method is employed for estimating the required probabilities.

The important issue when using histogramming to estimate probability density of a dataset, is the choice of bin width i.e. number of bins. In this work we used Doane's rule for bin number calculation [12]. Thus, having $N$ samples in the dataset, the number of bins is $J = 1 + log_2(N) + log_2(1 + \hat{k}\sqrt{N/6})$ where $\hat{k}$ stands for the standardized skewness coefficient. By using this rule for the number of bins, there is no a priori assumption whether the distribution of data is Gaussian or not, and $\hat{k}$ reflects the departure from normality.

Knowing the number of equally spaced histogram bins $j = 1,\ldots,J$ it is feasible to estimate the probability density of the feature vector. In our case the continuous random variable $Y$ is represented by the set of values obtained after projecting each image on the eigenvector space. Hence,

$$p(y) \approx \frac{n_j}{N}. \quad (5)$$

where $n_j$ stands for the number of observations in a histogram bin $j$, and $N$ is the total number of elements in a training set.

Similarly, if we have $n_c$ samples for each speech class, the class prior probabilities can be calculated from the data sample as:

$$p(c) = \frac{n_c}{N}, \quad N = \sum_{c=1}^{N_c} n_c. \quad (6)$$

In this work isolated-word speech recognition is considered and four different word ("digit") classes are available. Although lip movement vary along the different word pronunciations, there are some distinct lip shapes particular to each digit. Moreover, since we assume the audio and visual modalities are correlated and synchronized, each observed lip image sequence can be labeled with respect to the corresponding audio class. Thus, it is possible to estimate the conditional probability of the feature vector knowing the class, such that:

$$p(y|c) \approx \frac{n_{j|c}}{n_j}. \quad (7)$$

where $n_{j|c}$ is the number of features in the bin $j$ belonging to class $c$, while $n_j$ denote the total number of features in the corresponding bin.

After calculating mutual information for each principal component and word classes, eigenvectors are sorted on decreasing order of the mutual information values. The desired number of the most informative eigenvectors is used to find projections of each image from the database on the eigenspace. The obtained projection coefficients are employed as inputs of the visual feature vector for the task of isolated word recognition. Similarly to *eigenlips* where the

Figure 2: The first three mutual information principal components from the Tulips1 database.

Figure 3: Lip image examples from the Tulips1 database, [13]. *Upper row*: Original lip images. *Lower row*: Corresponding normalized lip images.

projection coefficients are chosen due to the maximum eigenvalues, these projections can be viewed as *mutual information eigenlips*. The first three mutual information principal components are shown in Figure 2. If we compare these images to those shown in Figure 1, it can be seen that MI principal components emphasize the lip shapes (edges) and discard the lighting directions.

## 3. EXPERIMENTAL FRAMEWORK

### 3.1 Audio-visual database

The work here utilizes material from the Tulips1 audio-visual database [13]. It is a small, publicly available database of 12 subjects, pronouncing the first four digits in English two times in repetition. The audio part is sampled at 11127 Hz with 8 bits per sample. The video part consists of 934 gray scale lip images of size $100 \times 75$, sampled at the rate of 30 fps.

### 3.2 Visual features

In order to account for possible head rotations, as well as for translation and scaling, the images from the database are further normalized. This task is performed in a similar way as in [14] from the manually marked images. Some original example images from the database, as well as the normalized ones are represented in Figure 3.

Besides the original raw images, their first order time derivatives i.e. delta images are taken into account. PCA and MI PCA are also applied on this image set. Each image from the dataset is projected onto the eigenfeature space and a certain number of projection coefficients is retained and used as a visual input for the recognition.

Furthermore, in order to have invariance to unwanted variations, a mean-removal PCA (MRPCA) as given in [9] is considered. The method is similar to Cepstral Mean Normalization [15]. Given the temporal mouth image sequence $\mathbf{x_1}, \ldots, \mathbf{x_T}$, new mouth images $\hat{\mathbf{x}}_\mathbf{t}$ are created by subtracting the mean image $\bar{\mathbf{x}} = \frac{1}{T}\sum_t \mathbf{x_t}$. PCA applied on this kind of images is regarded as mean removal MRPCA and in our case when using mutual information, denoted as mutual information mean removal PCA (MI MRPCA). Delta images are also found after applying mean removal on raw mouth images.

### 3.3 Acoustic features

Regarding the audio front end, the features of interest are commonly used Mel Frequency Cepstral Coefficients (MFCCs). These cues were calculated using HMM Toolkit (HTK) [15]. In order to have the audio features at the same rate as the visual ones the frame period is set to 30 Hz and features are extracted from the overlapping Hamming window of 50 ms duration. After processing of speech input, 12 MFFCs, energy term and their first and second order derivatives were extracted for each audio frame, resulting in a 39 dimensional feature vector.

After the audio and visual features are obtained, the corresponding feature vectors are provided as input to the Hidden Markov Model (HMM) based recognizer, built using the HTK Toolkit 3.2 [15]. Whole-word HMMs are used, one for representing each word class. Due to the small database/vocabulary size, the sub-word recognition was not feasible.

### 3.4 Results

Due to the limited size of the used database all experiments were done using a "leave-one-out" strategy. In each run 11 subjects were used for training and the remaining 12th was used for testing. The same procedure was repeated for each speaker and the results were averaged.

Audio-only recognition experiments were performed for both clean and noisy speech. Noisy environment was simulated by adding white Gaussian noise at different SNRs from 30 dB to $-18$ dB in steps of 6 dB. Training was done using clean data, while for testing purposes noisy speech samples were utilized. Several different system configurations were tested by changing the number of HMM states from two to five and the number of Gaussian mixtures from one to five. The best performance was achieved using a 3-state model with three mixtures per state and the obtained word accuracy rates are presented in Figure 4.

Since the main objective of this work is to show the benefits of using adequate visual cues, various visual-only recognition tasks were conducted on available images from the Tulips1 database. For the first set of experiments, traditional PCA and MI PCA were applied on raw gray scale images and the results are presented in Table 1.

The number of principal components retained for both methods was varied between 5 and 100. The best visual-only recognition results were obtained using HMMs with 5 or 6 states with one Gaussian per state and the highest rate was achieved for 35 MI PCA features. It is important to notice that no matter what the chosen number of features is, the mutual information based method performs better than ordinary PCA. This improvement is up to 10%. Furthermore, in order to capture the dynamics of lip movements besides

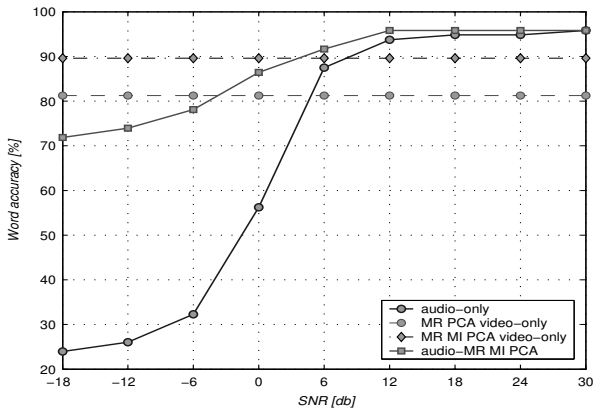


Figure 4: Audio-only, visual-only and audio-visual recognition results.

| | word accuracy [%] | |
|---|---|---|
| **# of features** | original PCA | MI PCA |
| 5 | 65.62 | 68.75 |
| 10 | 67.71 | 72.91 |
| 15 | 70.83 | 72.91 |
| 20 | 69.79 | 70.83 |
| 25 | 69.79 | 73.95 |
| 30 | 69.79 | 75.00 |
| 35 | 72.91 | 79.17 |
| 40 | 72.91 | 73.95 |
| 50 | 70.83 | 77.08 |
| 60 | 66.67 | 68.75 |
| 80 | 65.62 | 66.67 |
| 100 | 64.58 | 66.67 |

Table 1: PCA and MI PCA of original images. Visual-only recognition rates.

the original raw images we considered their first order time derivatives i.e. delta images. The same visual feature selection techniques were applied on such set. The inclusion of delta image features yielded to improved performance in both cases.

Table 2 shows the word recognition accuracy obtained using mean-removal PCA and mutual information mean-removal PCA, when changing the number of retained principal components from 10 to 100. It can be seen that the proposed method for feature selection using mutual information criteria clearly outperforms the traditional PCA based approach. The best results for visual-only recognition were achieved when the number of features was equal to 10 using mutual information principal components. Moreover, both the mean-removal MRPCA and mean-removal MI MRPCA perform better than those applied to the raw images. Taking into account delta features, the best achieved visual-only recognition rates are 89.6% for MI MRPCA, and 81.25% for MRPCA (see Figure 4). These rates were constant over the whole range of SNRs.

Finally, after testing various visual features our goal was to determine the accuracy of the audio-visual system when using both modalities. The integration of two modalities has been done utilizing feature fusion i.e. an early integration approach. The composite feature vector was obtained by

| | word accuracy [%] | |
|---|---|---|
| **# of features** | MRPCA | MI MRPCA |
| 10 | 80.21 | 87.50 |
| 15 | 81.25 | 84.37 |
| 20 | 81.25 | 86.46 |
| 25 | 81.25 | 84.37 |
| 30 | 79.17 | 85.42 |
| 35 | 75.00 | 83.33 |
| 40 | 73.95 | 75.00 |
| 50 | 70.83 | 75.00 |
| 60 | 68.75 | 71.87 |
| 80 | 66.67 | 70.83 |
| 100 | 65.62 | 66.67 |

Table 2: Mean-removal PCA and MI PCA of original images. Visual-only recognition rates.

simple concatenation of audio and visual cues. The training procedure was performed using joint feature vector of clean speech samples and mean-removal MI PCA features, while testing was done on noise-corrupted data. Figure 4 shows the obtained results using 39 acoustic features together with 15 MI MRPCA eigenfeatures from the original images and 15 features from the corresponding delta images, or equivalently a 69 dimensional feature vector. The HMM models used had 4-states with three Gaussian mixtures. Clearly, an audio-visual approach outperforms audio-only recognition when noise is present in the scene, as well as visual-only recognition. The improvement in word accuracy rate ranges from around 30% when the SNR equals 0 dB to around 45% at $-18$ dB. However, at low SNRs e.g. less than 0 dB the accuracy drops bellow the visual-only rate. The reason is the chosen feature fusion approach that reflects mismatch between visual and noise corrupted audio data.

## 4. CONCLUSION

We presented a novel approach for selecting the most informative eigenlips using mutual information criteria. Various visual-only isolated word recognition experiments were done using these features and recognition results show clearly that our method outperforms the conventional based one. Moreover, the recognition accuracy can be further improved by using the mean removal on an image sequence to reduce unwanted variations. Regarding the multi-modal approach, audio-visual recognition rates when using maximum mutual information eigenfeatures are higher than those of audio-only when the noise is present in the scene. Overall rates are also higher than visual-only in all cases but for low SNRs (less than 0 dB). The drop in accuracy rate is due to the chosen feature fusion method.

Future work would be to test the proposed method on a larger dataset and under the presence of different noise types. Also, other more advanced fusion strategies should be employed in order to overcome the problem of lower accuracy rates at low SNR levels, as well as to take into account visual anticipation/retention phenomena [16]. Since the mutual information criterion highly depends on the correct probability density estimation, one possible research direction is to utilize other density estimation techniques. Another important issue to consider is the extension of the proposed method to joint mutual information estimation to reduce the possible

redundancy in the selected eigenfeature set.

## REFERENCES


[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview," *Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.

[2] M. Kirby, F. Weisser, and G. Dangelmayr, "A model problem in the representation of digital image sequences," *Pattern Recognition*, vol. 26(1), pp. 63–73, 1993.

[3] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Proc. of ICASSP94*, Adelaide, Australia, April 19-22. 1994, pp. 669–672.

[4] G. Potamianos, H. P. Graf, and E. Cosatto, "An Image Transform Approach for HMM based Automatic Lipreading," in *Proc. of ICIP 1998*, Chicago, Illinois, October 4-7. 1998, pp. 173–177.

[5] H. H. Yang, S. Van Vuuren, S. Sharma, and H. Hermansky, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Communication*, vol. 31, pp. 35–50, 2001.

[6] D. Ellis and J. Bilmes, "Using mutual information to design feature combinations," in *Proc. of ICSLP 2000*, China, October 16-20. 2000, pp. 79–82.

[7] P. Scanlon, G. Potamianos, V. Libal, and S. M. Chu, "Mutual Information Based Visual Feature Selection for Lipreading," in *Proc. of ICSLP 2004*, South Korea, October 4-8. 2004, pp. 2037–2040.

[8] M. S. Gray, T. J. Sejnowski, and J. R. Movellan, "A Comparison of Image Processing Techniques for Visual Speech Recognition Applications," *Advances in Neural Information Processing Systems*, vol. 13, pp. 939–945, 2001.

[9] S. Lucey, "An evaluation of visual speech features for the tasks of speech and speaker recognition," in *Proc of AVBPA*, Guildford, U.K., 2003, pp. 260–267.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[11] R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*. New York: MIT Press & John Wiley & Sons, Inc. 1961.

[12] D. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: John Willey & Sons, 1992.

[13] J. R. Movellan, "Visual Speech Recognition with Stochastic Networks," *Advances in Neural Information Processing Systems*, vol. 7, pp. 851–858, 1995.

[14] J. Luettin, *Visual Speech and Speaker Recognition*, PhD Thesis, University of Sheffield, 1997.

[15] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book version 3.2*. Cambridge University Engineering Department, 2002.

[16] P. Deléglise, A. Rogozan, and M. Alissali, "Asynchronous Integration of Audio and Visual Sources in Bi-modal Automatic Speech Recognition," in *Proc. EUSIPCO 1996*, Trieste, Italy, September 10-13, 1996.