

Mutual Information Based Visual Feature Selection for Lipreading

Patricia Scanlon*, Gerasimos Potamianos, Vit Libal, Stephen M. Chu

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

{pscanlo, gpotam, libalvit, schu}@us.ibm.com

Abstract

Image transforms, such as the discrete cosine, are widely used to extract visual features from the speaker's mouth region to be used in automatic speechreading and audio-visual speech recognition. Typically, the spatial frequency components with the highest energy in the transform space are retained for recognition. This paper proposes an alternative technique for selecting such features, by utilizing the mutual information criterion instead. Mutual information between each individual spatial frequency component and the speech classes of interest is employed as a measure of its appropriateness for speech classification. The highest mutual information components are then selected as visual speech features. Extensions to this scheme by using joint mutual information between candidate feature pairs and classes are also considered. The algorithm is tested on visual-only speech recognition of connected-digit strings, using an appropriate audio-visual database. For low-dimensional visual feature vectors, the proposed method significantly outperforms features selected by means of energy, reducing word error rate by as much as 20% relative. These gains however diminish as higher feature dimensionalities are allowed.

1. Introduction

Visual speech information, extracted from the speaker's mouth region, has been repeatedly demonstrated to improve the performance and noise robustness of *automatic speech recognition* (ASR) [1–5]. Critical however to the performance of the resulting *audio-visual ASR* system is the choice of visual features that contain sufficient information about the uttered speech.

A popular approach to extracting such features is to use appearance-based techniques. These methods consider all pixels of the *region-of-interest* (ROI) around the speaker's mouth as potentially informative about visual speech, and they seek linear transforms of them in order to reduce feature vector dimensionality before classification, while retaining most relevant speech information. Popular image transforms employed in this framework are the principal component analysis [2] and the *discrete cosine transform* (DCT) [3–6]. The latter is very widely used because it is amenable to fast computations, while it also avoids expensive training: Typically, the DCT coefficients with the highest energy over a training dataset are used for visual speech recognition [4, 5]. However, while such energy-based criterion is appropriate for image compression, it does not necessarily provide high speech information content, suitable for discriminating between the speech classes of interest. On the other hand, a more suitable transform for this task, namely the *linear discriminant analysis* (LDA) [7], becomes computationally intractable as the ROI size becomes large (such as, for ex-

ample, a typically-sized 64×64 -pixel ROI).

In this paper, we propose using the *mutual information* (MI) criterion as a means of selecting the most speech informative features within the candidate pool of the DCT coefficients of the visual ROI. In particular, the MI between each individual spatial frequency component and the speech classes of interest is used as a measure of its appropriateness for speech classification. The highest mutual information components are then selected as visual speech features. Extensions to this scheme by using *joint mutual information* (JMI) between candidate feature pairs and classes are also considered. The proposed schemes are evaluated on visual-only speech recognition of connected-digit strings, using an in-house large audio-visual database, and are compared with our baseline visual-front-end, that uses energy-based DCT coefficient selection [5]. To our knowledge, this is the first attempt to select individual discriminative visual features from a pool of image transform coefficients. In particular, with respect to DCT features, only energy and variance have been considered for this task in the literature [4, 5], with the former being in general preferable [4].

Our work draws on related research previously applied to audio-only feature selection. For example, the MI and JMI criteria have been used for feature selection across time and frequency for both phone and speaker/channel classification [8]. Furthermore, in [9], MI has been calculated separately for each broad phonetic class to reveal clearer time-frequency structure. Vowel classification experiments performed showed that selecting input features based on the MI criteria provided a significant increase in accuracy.

The paper is structured as follows: Section 2 reviews necessary background on MI, followed by implementation details of the selection algorithm and the MI computation, as well as its JMI extension. Section 3 provides our experimental framework and results, followed by a summary in Section 4.

2. Mutual Information for Visual Feature Selection

2.1. Background and Notation

Let us denote by C the discrete random variable of speech class $c \in \mathcal{C}$, where, for example, set \mathcal{C} can consist of all sub-phonetic states used in ASR by means of a *hidden Markov model* (HMM). The *entropy* of C is defined as

$$H(C) = - \sum_{c \in \mathcal{C}} p(c) \log p(c), \quad (1)$$

and it represents a measure of uncertainty about its value [10].

One wishes to reduce this uncertainty, by observing appropriate features that convey speech information. In our case, these features are the DCT coefficients of the mouth ROI. We consider them as single-dimensional random variables $X \in \mathcal{X}$, where \mathcal{X} represents the entire ROI in the transform domain,

*Work performed while on internship at the IBM T.J. Watson Research Center. Patricia Scanlon is with the Electrical and Computer Engineering Dept., University College Dublin, Dublin 4, Ireland; email: patricias@ee.ucd.ie

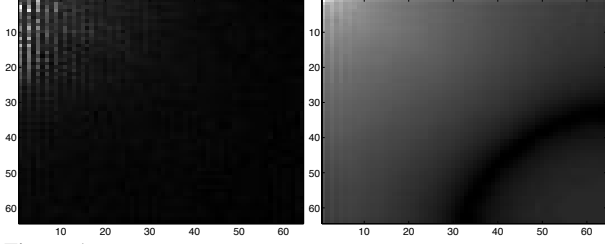


Figure 1: *Mutual information (left) and energy plots (right) for all 64×64 DCT coefficients.*

for example all 64×64 transform components (assuming a two-dimensional DCT on a 64×64 -pixel ROI, as is the case in our system described in Section 3.1). Random variables X take on real values $x \in \mathcal{R}$. Observation of visual feature X results in reduced uncertainty about C , measured by the mutual information $I(C; X)$, given by [10]

$$I(C; X) = H(C) - H(C|X) = H(X) - H(X|C). \quad (2)$$

Note that $I(C; X) = I(X; C)$, and that in general, $0 \leq I(X; C) \leq \min\{H(X), H(C)\}$. Furthermore, $I(X; C) = 0$, if and only if X and C are independent. Using (1) and (2) (see also [10]), one can obtain the mutual information equation

$$I(X; C) = -\int_{x \in \mathcal{R}} p(x) \log p(x) + \sum_{c \in \mathcal{C}} p(c) \int_{x \in \mathcal{R}} p(x|c) \log p(x|c). \quad (3)$$

From the above, a clear strategy for selecting features $X \in \mathcal{X}$ is to seek the ones that maximize (3). The resulting algorithm and its implementation are described next.

2.2. The Selection Algorithm and its Implementation

Putting aside for the moment the issue of computing (3), the MI-based algorithm for visual feature selection within the candidate pool of the DCT coefficients of the visual ROI can be easily expressed as:

$$X_i = \arg \max_{X \in \mathcal{X} - \mathcal{X}_{i-1}} \{I(X; C)\}, \text{ and } \mathcal{X}_i = \mathcal{X}_{i-1} \cup X_i, \quad (4)$$

for $i = 1, 2, \dots, d$, with $\mathcal{X}_0 = \emptyset$, where d is the desired dimensionality of the selected feature vector. Note that this approach represents a simple sorting of all mutual information values (for example, 4096 of them for a 64×64 -pixel ROI), and it results in a nested selected coefficient set $\mathcal{X}_1 \subset \dots \subset \mathcal{X}_d \subset \mathcal{X}$.

To obtain estimates of the MI values, needed in (4), we follow the *histogram approach* for approximating the density functions required in (3), as in [8, 11]. In this approach, we first decide on the number of histogram bins to be used. Following [8], we consider Doane's rule and approximately use $K = \log_2 n + 1 + \log_2(1 + \hat{k}\sqrt{n/6})$ bins to estimate $p(X|C)$ and $P(X)$. In the above-mentioned rule, \hat{k} is the estimate of the kurtosis of the DCT coefficient of interest (i.e., of random variable X), and n is the total number of training samples. In our experiments, $n \approx 100k$, and, on the average, 30 bins are derived for each DCT coefficient. Note that the kurtosis estimates indicate that our DCT coefficient data are strongly non-Gaussian.

Given the number of bins, we form equally spaced intervals b_k , $k = 1, 2, \dots, K$, between the minimum and maximum observed values of each X in the training data. Then we approximate $p(x) \approx n_k/n$, iff $x \in b_k$, where n_k denotes the number of observations $x \in b_k$. Assuming that class labels $c \in \mathcal{C}$ are available for the training samples, we can

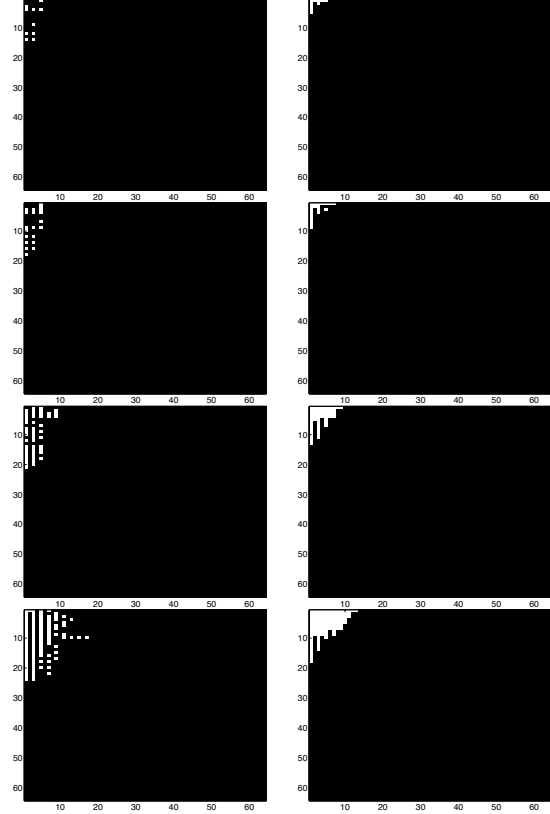


Figure 2: *Mutual information (left) and energy (right) based selection masks for feature dimensionality of $d=10,20,50,100$ (top to bottom).*

similarly also obtain the n_c and $n_{k,c}$ counts, thus estimating $p(c) = n_c/n$ and approximating $p(x|c) \approx n_{k,c}/n_c$, for all $x \in b_k$, $k = 1, 2, \dots, K$, and $c \in \mathcal{C}$. Based on these estimates, computation of (3) becomes feasible.

In our experiments, the required class labels are obtained by utilizing the bimodality of the data, i.e., the fact that time-synchronous acoustic observations are also available in addition to the visual ones. In particular, the visual feature vectors (DCT coefficients) are linearly interpolated from 30 Hz (video capture rate) to the audio feature extraction rate of 100 Hz used in our audio-visual ASR system [5] (see also Section 3.1). Thus, a *forced alignment* in the acoustic modality, using the data speech transcriptions and a trained audio-only HMM, can provide class labels for all audio feature vectors, and thus for their time-synchronous DCT features. Such classes can be the context-dependent HMM states, or, at a coarser level, their corresponding phones.

We now present some examples of the practical application of this scheme on our data (described in Section 3.2), as opposed to the baseline algorithm of DCT feature selection based on energy [5]. Figure 1 shows the MI vs. the energy values for all 64×64 DCT coefficients, computed over our training data. Notice that the energy plot shows a gradual increase in energy from the DC component (top left corner), to the highest frequency components (bottom right corner), but with slightly higher energy in the odd columns. It also appears that features in the odd columns have much higher MI than those in the even ones, a fact that may be related to the lateral face symmetry.

Feature selection masks are subsequently generated using (4) for $d = 10, 20, 50, 100$. Figure 2 shows the four MI masks, and its four corresponding energy-based masks, with the same



Figure 3: Original images (left), reconstructed using the MI mask - including the DC coefficient - (center), and reconstructed using the energy mask (right). Both reconstructions are with $d=20$ features selected.

feature vector dimension. Interestingly, the energy masks maintain the same pattern as their dimension grows, with emphasis on the odd column components. The MI masks have less of a consistent pattern than the energy ones, especially in the lower dimensions. It can also be seen in the MI masks that information is spread out over numerous non-contiguous spatial frequency components that incorporate more higher frequency components than the energy masks.

In order to illustrate the differences in the masks, Figure 3 shows three original images reconstructed using the MI and energy masks with $d=20$. While in both sets of images there is a relatively low level of detail compared to the original image, the reconstruction using the MI mask maintains a very clear outline of the mouth region, while the mouth is extremely blurred using the energy mask. Note also, that in low dimensions, e.g. $d = 10, 20, 50$, the DC coefficient is not selected for the MI mask, in fact it is ranked as the 78th most important feature based on MI. In contrast, in the energy mask, it is ranked first. To make the images visible on reconstruction, the DC coefficient for the MI mask is activated, i.e. only 19 features plus the DC coefficient are used for the MI mask reconstruction.

2.3. Joint Mutual Information for Feature Selection

As the MI for each feature is computed separately, there may be some redundancy in the features selected using the MI criterion. In [11], a modified MI-based feature selection algorithm is proposed that proceeds in a “greedy” selection of features. The algorithm first obtains the feature with the highest MI, and then it searches the remaining features for a candidate that maximizes a combination of MI with respect to the classes and the MI with respect to each of the previously selected features. Namely, for the selection of the i^{th} feature, the scheme computes quantity

$$I(X; C) - \beta \sum_{X' \in \mathcal{X}_{i-1}} I(X; X'), \quad (5)$$

for all $X \in \mathcal{X} - \mathcal{X}_{i-1}$, i.e., it replaces $I(X; C)$ by (5) in (4). In (5), β is an appropriately chosen quantity between 0 and 1 [11].

Notice that this method is suboptimal as it does not include information on the joint MI (JMI) between *all* preselected and each candidate feature with the class labels. However, the data

required to reliably estimate the JMI increases exponentially with the number of preselected features, and hence this is not a practical approach. Thus, we only extend (5) to include JMI between two features and the class random variables. In particular, our first scheme replaces in (4) quantity $I(X; C)$ by the maximum JMI of X and each of the already selected features with respect to the classes, namely by

$$\max_{X' \in \mathcal{X}_{i-1}} I(X, X'; C) = I(X; C) + \max_{X' \in \mathcal{X}_{i-1}} I(X; C|X') \quad (6)$$

(see also [10]). A second proposed scheme uses

$$I(X; C) + \frac{1}{|\mathcal{X}_{i-1}|} \sum_{X' \in \mathcal{X}_{i-1}} I(X; C|X') \quad (7)$$

instead. The right most quantity can be also computed as a function of JMI $I(X, X'; C)$, using a two-dimensional histogram method for estimating its two entropy terms:

$$I(X, X'; C) = H(X, X') - H(X, X'|C),$$

similarly to what was discussed in Section 2.1.

3. Experiments

We now proceed to investigate the performance of the proposed MI-based feature selection algorithms compared to the baseline energy-based selection scheme. For this purpose, we conduct a number of visual-only recognition experiments on a connected-digits database. Before reporting the results, we briefly describe the visual-front-end processing of our system, the visual ASR approach, and the audio-visual database.

3.1. The Visual Processing and Recognition System

Given full-face video of the speaker’s face, our system first utilizes a statistical face and facial feature detector to estimate the location of the face and of landmark facial features. Given these, a greyscale 64×64 -pixel ROI is extracted, normalized for lighting, as well as for head size and orientation. A two-dimensional DCT is then applied to the ROI, resulting in a 4096-dimensional vector of transform coefficients. Our baseline system selects 100 of these coefficients with the highest energy. Here, this selection scheme is replaced by a number of MI-based algorithms. Subsequently, the selected coefficients are mean-normalized and interpolated to 100 Hz. To reduce their dimensionality in a discriminant way, an optional cascade of an LDA projection followed by a maximum likelihood linear transform (MLLT) rotation is applied to the features. This is referred to as the *intra-frame* LDA/MLLT, and results into 30-dimensional features. Finally, in order to incorporate temporal information, a second, *inter-frame* LDA/MLLT is applied on a concatenation of 11 consecutive feature frames, resulting to a final visual feature vector of dimension 41. Such approach is usually preferable to the augmentation of the static features by their first and second derivatives. More details on the visual-front-end can be found in [5].

Once visual features are extracted, they are provided to an HMM-based recognizer. The HMMs are 3-state left-to-right phone models, consisting of context-dependent sub-phonetic states, with their parameters estimated by the Expectation-Maximization algorithm. For the particular recognition task at hand (connected-digit strings), the HMMs correspond to 22 phones, 66 sub-phonetic states (3 states per phone), 159 context-dependent states, and 3.2k Gaussian mixture components. Decoding is performed using a stack decoder with an 11-word vocabulary (digits 0-9 and “oh”) and no grammar.

Table 1: Test-set visual-only WER, %, using various DCT feature selection schemes. In the center column, 100 features are selected, and their dimensionality reduced by intra-frame LDA/MLLT. In the right column the first 5 only features are chosen. In both cases, inter-frame LDA/MLLT is also applied to obtain final, 41-dimensional features.

Feature dimension d	100	5
Intra-frame LDA/MLLT	Yes	No
Energy	28.56	75.69
MI; 22 classes; uniform class priors	32.24	65.48
MI; 66 classes; uniform class priors	30.34	65.27
MI; 66 classes; eq. (4)	28.88	59.69
JMI-max; 66 classes; eq. (6)	29.46	60.65
JMI-avg; 66 classes; eq. (7)	30.40	58.79

3.2. The Audio-Visual Database

The database considered in our experiments contains high-quality frontal full-face video of 50 subjects uttering connected-digit strings. The corpus is recorded in a quiet studio-like environment using a high-quality camera, uniform background and lighting, and relatively stable frontal subject pose. The video is MPEG2-encoded at a resolution of 704x480 pixels, processed at 30 Hz. Wideband audio is also available at about 20 dB signal-to-noise ratio [5]. From these data, close to 8 hrs are used for training, and about 55 minutes for testing (623 sequences), assuming a multi-speaker training/testing paradigm. For the particular test-set size, a difference of about 1.6% in word error rate (WER) is significant at the 5% level.

3.3. Results and Discussion

To verify the hypothesis that high MI DCT locations correspond to relevant features for discrimination, several visual speech recognition experiments are performed. The table provides WERs, %, comparing the baseline method to several MI approaches, both when $d=5$ only static features are selected with no intra-frame LDA/MLLT, as well as when $d=100$ features are chosen followed by the intra-frame dimensionality reduction by LDA/MLLT. In both cases, a temporal, inter-frame LDA/MLLT is applied over 11 frames to obtain final, 41-dimensional visual features. Results show a 20% relative improvement achieved by the MI method over the baseline, when using the algorithm of (4) to select the top 5 features and a set \mathcal{C} of 66 classes (75.69% \rightarrow 59.69%). However, these gains disappear when all 100 features are selected (the difference between them is statistically insignificant).

In addition, from the table, it can also be seen that using 66 classes provides a significant improvement over 22 classes in the computation of the MI mask. A further improvement is seen when using the priors of the classes over using uniform priors $p(c)$ in the computation of MI (3). Employing JMI as in (6) and (7) did not significantly change results, although using average JMI seems preferable to the maximum JMI for the case where the first 5 only features are utilized.

Finally, in Figure 4, we further compare the baseline scheme with the best MI selection technique of the table (66-class MI, by (4)). In particular, we are interested to see how the two algorithms compare over a wide range of chosen feature dimensionalities. It is clear that MI-based feature selection consistently outperforms the energy-based method for up to $d=25$, for both approaches used to incorporate temporal information (inter-frame LDA/MLLT or derivatives). Above $d=30$ though, any differences become statistically insignificant, with no improvement obtained from the MI method over the baseline. This may be due to the fact that for $d=30$ and above, the most relevant features for class discrimination are already included in the energy feature mask.

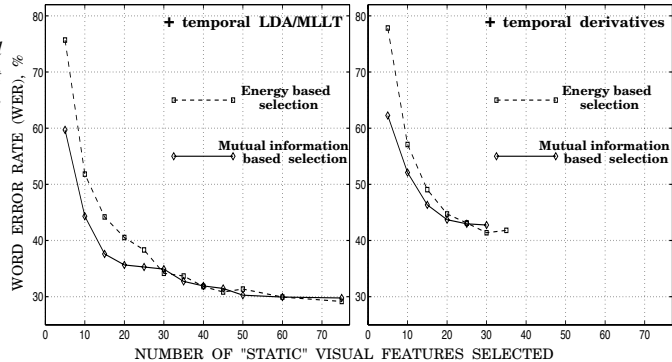


Figure 4: Visual-only WER as a function of the number d of selected coefficients using energy vs. MI by means of (4). Two approaches are used to incorporate temporal information. Left: An inter-frame LDA/MLLT (final features are 41-dimensional); Right: Inclusion of first- and second-order derivatives (final features have dimension $3d$).

4. Summary

We investigated alternative techniques to the energy-based selection of DCT visual speech features, utilizing the mutual information criterion. The hypothesis proposed that features with high mutual information are more discriminative for visual speech recognition than baseline features chosen based on the highest energy was verified for low-dimensional feature vectors, where the proposed algorithm significantly outperformed the baseline by reducing visual-only word error rate by as much as 20% relative. These gains however diminished as higher feature dimensionalities were allowed.

5. References

- [1] E.D. Petajan, "Automatic lipreading to enhance speech recognition," *Proc. Global Telecomm. Conf.*, pp. 265–272, 1984.
- [2] C. Bregler and Y. Konig, "'Eigenlips" for Robust Speech Recognition," *Proc. Int. Conf. on Acoustics, Speech, and Signal Process.*, pp. 669–672, 1994.
- [3] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," *Proc. Int. Conf. Spoken Lang. Process.*, pp. 547–550, 1994.
- [4] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," *Proc. Int. Conf. Spoken Lang. Process.*, pp. 1925–1928, 2002.
- [5] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, 91(9): 1306–1326, 2003.
- [6] P. Scanlon and R. Reilly, "Visual feature analysis for automatic speechreading," *Proc. Work. Audio Visual Process.*, pp. 127–132, 2003.
- [7] G. Potamianos and H.P. Graf, "Linear discriminant analysis for speechreading," *Proc. Work. Multimedia Signal Process.*, pp. 221–226, 1998.
- [8] H.H. Yang, S. Van Vuuren, S. Sharma, and H. Hermansky, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Communication*, 31: 35–50, 2000.
- [9] P. Scanlon, D.P.W. Ellis, and R. Reilly, "Using mutual information to design class specific phone recognizers," *Proc. Europ. Conf. Speech Technol.*, pp. 857–860, 2003.
- [10] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.
- [11] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, 5(4): 537–550, 1991.