Information Theoretic Feature Extraction for Audio-Visual Speech Recognition

Mihai Gurban, *Member, IEEE*, and Jean-Philippe Thiran, *Senior Member, IEEE*

***Abstract*—The problem of feature selection has been thoroughly analyzed in the context of pattern classification, with the purpose of avoiding the curse of dimensionality. However, in the context of multimodal signal processing, this problem has been studied less. Our approach to feature extraction is based on information theory, with an application on multimodal classification, in particular audio–visual speech recognition. Contrary to previous work in information theoretic feature selection applied to multimodal signals, our proposed methods penalize features for their redundancy, achieving more compact feature sets and better performance. We propose two greedy selection algorithms, one that penalizes a proportion of feature redundancy, while the other uses conditional mutual information as an evaluation measure, for the selection of visual features for audio–visual speech recognition. Our features perform better than linear discriminant analysis, the most usual transform for dimensionality reduction in the field, across a wide range of dimensionality values and combined with audio at different quality levels.**

***Index Terms*—Audio–visual speech recognition, feature selection, mutual information.**

I. Introduction

MULTIMODAL signal processing analyzes a physical phenomenon through several types of measures, or modalities. This leads to the extraction of higher-quality and more reliable information than that obtained from single-modality signals. The advantage is two-fold. First, as the modalities are usually complementary, the end-result of multimodal processing is more informative than for each of the modalities individually. This is true in all application domains: human-machine interaction, multimodal identification or multimodal image processing. The second advantage is that, as modalities are not always reliable, it is possible, when one modality becomes corrupted, to extract the missing information from the others, leading to a more reliable system.

To offer an example for the first advantage of multimodal systems, the complementarity of information, in multimodal medical image analysis [1], [2], information that is missing in one modality may be clearly visible in another. The same is true for remote sensing [3]. Here, by missing information we mean information that exists in the physical reality but was not captured through one of the particular modalities used. By using several different modalities we can get closer to the underlying phenomenon, which might not be possible to capture with just one type of sensor.

Another example could be audio–visual speech recognition (AVSR), which is a method of improving automatic speech recognition results through the use of information from the visual modality (the motion of the speaker's lips) [4]. This works particularly well when the audio modality is corrupted by noise, but it can also bring a slight improvement when the audio is clean. The reason is that some phonemes are more easily distinguishable in video than in audio, and humans themselves use this subconsciously, as proven by the McGurk effect [5].

For the second advantage, the improved reliability of multimodal systems, consider the case of a three-modal biometric identification system, based on face, speech and mouth motion [6]. When one of the modalities is unreliable, the system might still be able to identify the person correctly based on the other modalities. AVSR [4] can also be given as an example, since, if for some reason the video becomes unavailable, the system should seamlessly revert to audio-only speech recognition.

There are two essential challenges in multimodal signal processing. The first one is that features used from each modality need to be as relevant and as few as possible. The fact that multimodal systems have to process more than just one modality means that they can run into errors caused by the curse of dimensionality much more easily than mono-modal ones since the dimensionality of the combined multimodal signal is higher than each of the individual modalities. Originally, the term *curse of dimensionality* was introduced by Bellman to show that the number of points necessary to uniformly sample a volume of space grows exponentially with the dimensionality [7]. This has important implications in the classification domain, since accurate models can only be obtained if an adequate number of samples is available, and obviously this required number of samples grows with the dimensionality of the features.

Traditionally, the constraining assumption of joint Gaussianity was used to simplify the modeling of multimodal signals [8]–[10]. However, the results were unsatisfactory, since the relations between signals from different modalities is quite complex, as for example the dependency between mouth movement and sound for speech, so a Gaussian model is not appropriate. More complex models require a lower dimensionality or a much higher number of samples. To avoid this problem, the projection of multimodal data to low-dimensional subspaces has been proposed [11], although by performing a

Manuscript received December 11, 2008; accepted June 04, 2009. First published June 30, 2009; current version published November 18, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lucas Parra. This work is supported by the Swiss National Science Foundation through the IM2 NCCR.

The authors are with the Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne, 1015 Ecublens, Switzerland (e-mail: mihai.gurban@epfl.ch; jp.thiran@epfl.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2009.2026513

costly optimization of the mutual information between modalities in the transformed space.

Most machine learning models need to attribute some parameters to cover the variability in the input features. The more parameters, the higher the complexity or capacity of the classifier. If some of the features are just noise, the capacity allocated for them is practically wasted. It is also possible that false regularities are found in the unneeded features, also leading to a waste of capacity. Dimensionality reduction is thus a necessary step in any application which requires modeling complex signals, and is achieved through selection, transforms or the combination of the two.

The second essential challenge is multimodal integration. Since the signals involved do not necessary have the same rate, range or even dimensionality, combining information coming from such different sources is not straightforward. This can be done at different levels, starting from the basic signal level by combining the signals themselves, if they are compatible, up to the highest decision level, where only the individual decisions taken based on the signals are combined.

In this paper, we will focus on the first topic, dimensionality reduction for multimodal signals. Feature selection is important in multimodal signal processing, where the dimensionality of the data can get particularly high.

Our application is AVSR [4], where the visual modality only contributes in a small part to the final result, however, its dimensionality can surpass several times that of the audio stream. Extracting only the relevant information from the visual modality in a low-dimensional vector is then very important here, not only for a good recognition accuracy, but also when taking into account the efficiency of the system.

Our contribution is a method of selecting visual features and thus reducing the dimensionality of the visual feature vector for audio–visual speech recognition. The novelty of the presented work consists in the way a redundancy penalty was introduced in the measure used to select features in the particular context of AVSR. Our method is based on maximizing mutual information (MI) between the features and the class labels, while also minimizing their redundancy with respect to the same class labels. Although methods of penalizing redundancy for feature selection are well-known in the general classification domain, to our best knowledge, they were not applied before to the AVSR problem. The MI feature selection methods used in AVSR [12], [13] only maximize MI without penalizing for redundancy.

Another contribution is the fact that we present an extensive evaluation of our proposed feature selection methods based on MI, showing both monomodal and multimodal results, for a wide range of dimensionality values. Similar work in the literature typically uses visual features having a dimensionality around 40, which we prove could be too high for some applications. For example, in [14] the visual features dimensionality is 35, while in [15] it is 41. We present results for dimensionality values varying between 4 and 192 features, and show how, because of the curse of dimensionality, small feature sets can perform better than larger ones.

This paper continues and expands our previous work presented in [16].

II. Feature Selection for Classification

In this section we will present a short overview of feature selection for classification in general, while the application to AVSR will be detailed in the next section. A good overview of dimensionality reduction methods in the context of classification can be found in [17].

A. Feature Extraction

Feature extraction is the process of deriving useful information from an original signal, information that is relevant for the task and also has a more compact representation, suitable for use in a classifier. This can be achieved simply through *selection*, in which elements of the original data vector are kept, or through a *transform*, which will project the original data in a different, lower-dimensional space.

In *unsupervised* feature extraction, no class information is used, only statistical information of the features. On the other hand, if class information is included, the dimensionality reduction process is *supervised*.

The simplest supervised feature extraction methods are *filters*, where the criterion to assess the quality of a feature subset is some measure computed directly on the features. A second category are the *wrappers*, which use the final accuracy given by the classifier itself as a measure, requiring for this the complete training and testing of a classification system for each subset of features taken into consideration. This requires significantly more resources, but potentially will lead to feature sets which are better adjusted to the specific classifier used.

We will focus now on filters which perform feature selection, as they lead to a good balance between the quality of the obtained features and the resources, both time and computing power, which are required. Assume that we have a set F of n features, out of which we want to select a subset S of m features. The total number of possible subsets S, $\binom{n}{m}$, is very high, so processing every possible subset is typically impossible. The goal here is to obtain a subset S which retains as much from the information in F as possible. This could be achieved without being exhaustive, if the subset S is built iteratively with a greedy algorithm [18]. Such an algorithm works by selecting the "best" feature at each step, according to some measure of the quality of the features. In this way, a suboptimal solution is found, but hopefully one that is close to the overall optimum.

The quality measure used in the filter can be anything from statistical variance or correlation with the class labels to the mutual information (MI) between the feature and the labels. The search method can also vary, as there are other alternatives to sub-optimal greedy search. One example is the "branch-and-bound" algorithm [19], however, its effectiveness is limited by the requirement of monotonicity on the optimality measure.

Our method of choice is greedy search using MI as a feature quality measure. There are two reasons for this. The first comes from the definition of MI, as it can expose a dependency between two random variables, even when that dependency is nonlinear. This makes MI a powerful measure of dependency. The second reason is Fano's inequality [20], which can be interpreted in a way that shows that the bound on the probability of classification error can be lowered by choosing feature sets with higher MI with the class label.

B. Selection With Information-Theoretic Criteria

In general *mutual information* represents the reduction of uncertainty in one random variable when the value of another (related) random variable is known. In particular, if $I(Y;C)$ is the mutual information between a feature Y and the class labels C, it represents the amount of information gained about the class if the feature Y is used. A high mutual information here shows that the feature is relevant for the classification task and should be part of the subset of selected features.

The justification for using MI comes from Fano's inequality [20], written in this form:

$$p_e \geq \frac{H(C\,|\,F)-1}{\log N} = \frac{H(C)-I(C;F)-1}{\log N} \tag{1}$$

where N is the number of classes, F is the feature set and H is the entropy. Note that the expression contains the MI between the class labels and a whole set of features, not just one, as in the example above. The equation gives a lower bound for the probability of error, but does not guarantee that this lower bound will be reached by the classifier. It is clear that with a "bad" feature set, one which has a low mutual information with the class label, the bound on the probability of error will be high, forcing it to be high itself. However, when the bound is low, it is up to the actual classifier to come as close as possible to this bound. This shows that bad features lead to bad classification, but good features do not necessarily lead to good classification.

This shows that a feature set with a high mutual information with the class labels is desirable. However, computing mutual information from data is not trivial. The estimation of probability density functions is required, which is not practical in high dimensions, because it requires an unfeasible number of samples and a long training time. This is why most feature selection algorithms that use mutual information actually use two or three-dimensional measures [11]–[13], [21]–[25], never more. This means that at most two features are used together with the class label to compute the joint probability density.

We will present now a few information-theoretic feature selection methods. First, let us introduce the formal framework for these algorithms. Let $F = \{Y_1, Y_2 \ldots Y_n\}$ be the initial set of features. Let $\{\pi_1, \pi_2 \ldots \pi_m\}$ be a permutation on a subset of dimension m of the set of feature indexes $\{1 \ldots n\}$. Then the set of selected features can be written as $S = \{Y_{\pi_1}, Y_{\pi_2} \ldots Y_{\pi_m}\} \subset F$.

The simplest way to obtain a subset S iteratively would be to pick at each step the feature with the highest mutual information with the class labels. Formally, this means choosing at step $k+1$ the feature [17], [26]

$$Y_{\pi_{k+1}} = \underset{Y_i \in F \setminus S_k}{\arg\max}\, I(Y_i;C) \tag{2}$$

where $S_k = S_{k-1} \cup \{Y_{\pi_k}\}$ is the set of features selected at step k. This is equivalent with assuming that the mutual information that we want to maximize, $I(S;C)$, can be approximated with the sum of individually computed mutual information values $I(Y_k;C)$, with $Y_k \in S$.

However, this does not take into account any redundancy that may be present in the features. At the extreme, if two features have identical values and a high mutual information with the labels, they will both be chosen, even if the second feature does not bring any new information. So, in order to keep the set of relevant features small, redundancy should be penalized.

Redundancy between features can also be expressed in information-theoretic terms. Indeed, the redundancy between features Y_i and Y_j is measured by their mutual information, $I(Y_i;Y_j)$. However, as the set of selected features grows, we need to compute the redundancy of the candidate feature with the whole set of previously selected features, that is $I(Y_k;S_{k-1})$. This again requires high-dimensional probability density functions. The same approximation as for (2) can be applied, that is, $I(Y_k;S_{k-1})$ is the sum of individual mutual information values $I(Y_k;Y_i)$ with $Y_i \in S_{k-1}$. An algorithm that does just that is the MIFS algorithm [21]:

$$Y_{\pi_{k+1}} = \underset{Y_i \in F \setminus S_k}{\arg\max} \left[I(Y_i;C) - \beta \sum_{Y_{\pi_j} \in S_k} I\left(Y_i;Y_{\pi_j}\right) \right]. \tag{3}$$

Here the redundancy is approximated not with the sum, but with a proportion β of the sum, which the authors recommend setting to between 0.5 and 1.

A similar approach is to penalize the average redundancy [27]:

$$Y_{\pi_{k+1}} = \underset{Y_i \in F \setminus S_k}{\arg\max} \left[I(Y_i;C) - \frac{1}{|S_k|} \sum_{Y_{\pi_j} \in S_k} I\left(Y_i;Y_{\pi_j}\right) \right] \tag{4}$$

where $|S_k|$ is the size of set S_k. In the end, none of these methods has a good theoretical justification, since the high-dimensional mutual information values simply cannot be approximated with lower-dimensional ones.

Perhaps a little better justified theoretically are the information-theoretic methods based on the conditional mutual information, (CMI) as a measure [22], [25], $I(X;C\,|\,Y) = I(X,Y;C) - I(Y;C)$. This shows how much the random variable X increases the information we have about C when Y is given. The selection criterion is the following:

$$\begin{aligned} Y_{\pi_{k+1}} &= \underset{Y_i \in F \setminus S_k}{\arg\max} \left[\min_{Y_{\pi_j} \in S_k} I\left(Y_i;C\,|\,Y_{\pi_j}\right) \right] \\ &= \underset{Y_i \in F \setminus S_k}{\arg\max} \left[I(Y_i;C) - \max_{Y_{\pi_j} \in S_k} I\left(Y_i;Y_{\pi_j};C\right) \right] \end{aligned} \tag{5}$$

using $I(X;Y;C) = I(Y;C) - I(Y;C\,|\,X)$ [20]. The formula shows that, in fact, using CMI is also equivalent to penalizing redundancy, only it is redundancy of a different kind, $I(Y_i;Y_{\pi_j};C)$, which we could call *relevant redundancy*, since it also depends on the class. For a certain Y_i, the particular Y_{π_j} is found with which Y_i is most redundant, that is, which has the minimum conditional mutual information with the class label. By taking the maximum over this CMI, the feature that adds the most relevant information to this feature, and, implicitly, to the set S_k, is found.

This is also an approximation, since, as the set S_k grows, we should compute the conditional mutual information with respect

to the whole set, not its one most redundant feature. However this is impossible for the same reasons mentioned before.

In the end, the goal of all these algorithms is to maximize the joint MI between the S and C, which could be expanded like this (chain rule [20]):

$$\begin{aligned} I(S;C) &= I\left(Y_{\pi_1}, Y_{\pi_2}, \ldots, Y_{\pi_m}; C\right) \\ &= \sum_{k=1}^{m} I\left(Y_{\pi_k}; C \,|\, Y_{\pi_1}, \ldots, Y_{\pi_{k-1}}\right) \\ &= \sum_{k=1}^{m} \left[I\left(Y_{\pi_k}; C\right) - I\left(Y_{\pi_k}; C; Y_{\pi_1}, \ldots, Y_{\pi_{k-1}}\right)\right] \\ &= \sum_{k=1}^{m} \left[I\left(Y_{\pi_k}; C\right) - I\left(Y_{\pi_k}; C; S_{k-1}\right)\right]. \end{aligned} \tag{6}$$

An iterative algorithm could maximize the terms of this sum one by one

$$Y_{\pi_k} = \arg\max_{Y_i \in F \setminus S_k} \left[I(Y_i; C) - I(Y_i; C; S_{k-1})\right]. \tag{7}$$

Since Y_{π_k} is the particular Y_i that maximizes the kth term of the sum, all previously mentioned criteria (3), (4), and (5) can be interpreted as approximations of this general optimization. They all maximize the difference between $I(Y_i; C)$ and an approximation of the redundancy $I(Y_i; C; S_{k-1})$ between Y_i, S_{k-1} and the class labels C, which can be regarded as a penalty for redundancy. Several other algorithms follow this general path of maximizing the difference between $I(Y_i; C)$ and a redundancy penalty [23], [24]. However, nothing can be said about which of these approximation is actually better, since it all depends on the particularities of the high-dimensional probability density which cannot be estimated.

III. Feature Selection in AVSR

There are two major categories of visual features currently used for speech recognition. The first one is derived from image compression techniques, trying to represent the pixels of the region of the mouth with a compact feature vector. This is done through a transform, either the discrete cosine transform (DCT) [28], principal components analysis (PCA), linear discriminant analysis (LDA) [29], [30] or a cascade of these. The second type of visual features that are widely used are features derived from the extracted contour of the lips [14], [31]. However, due to variations in illumination, skin color, facial hair and so on, this contour is typically not robust enough for the task. We will mainly focus on transform-based visual features.

A typical AVSR system with image transform features would use DCT to compress the image [32], [33], keep only the low spatial frequencies [32] and then apply LDA on the resulting coefficients, reducing the dimensionality even more [4]. Although the classes do not have Gaussian probability density functions, which is a basic assumption for the LDA, the transform performs quite well. For example, a cascade application of LDA, first on each modality separately, and then on a concatenated multimodal feature vector, is the basis of the hierarchical LDA (HiLDA) transform [4].

An alternative to LDA would be to use MI as a selection criterion for DCT coefficients. Indeed, MI has been used before in both audio-only speech recognition and AVSR. For audio-only speech, both the maximum MI [34]–[36] and the CMI [37] methods have been used to assess the quality of audio features. The analysis in [34] shows that critical-band spectral energy observations are not Gaussian distributed, justifying the use of a non-linear measure such as the mutual information.

In AVSR, mutual information was used to select the relevant DCT coefficients from the visual modality. In [12] and [38], the authors select the features used for visual speech recognition based on either the mutual information between features and class labels, or the joint mutual information between two features and the class label. Formally, they use either

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} I(Y_i; C) \tag{8}$$

or

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} I(Y_i; C) + \frac{1}{|S_k|} \sum_{Y_j \in S_k} I(Y_i; C \,|\, Y_j) \tag{9}$$

where $|S_k|$ is the number of elements in S_k. The second term in (9) comes from the joint mutual information $I(Y_i, Y_j; C)$. Although proposed as an extension to MIFS [21], this algorithm does not include redundancy, as the emphasis in (9) is on $I(Y_i; C)$ with no penalty. Their findings show that the coefficients in the odd columns of the DCT have a much higher relevance, because of the symmetry of the mouth, as confirmed in [39]. We chose to exploit this property and use DCT coefficients from the odd columns only, as will be detailed in the following sections.

Another widely used transform for reducing the dimensionality of visual data is the PCA. Here, dimensionality reduction is typically achieved by only choosing the features which correspond to the largest eigenvalues. Selecting PCA coefficients from mouth images based on mutual information gives rise to "mutual information eigenlips" [13], leading to an improved speech recognition performance.

In the following, we will present our two proposed methods for visual feature selection in AVSR, both based on mutual information with a penalty for redundancy between the features in the chosen set. This approach is novel to the field of AVSR and leads to performance improvements compared to both the simple maximum MI method and to the LDA transform.

IV. Selecting Visual Features for AVSR With Mutual Information

In this section we will present our audio–visual speech recognition system, the features we used and the method of selection employed to reduce their dimensionality.

A. The Database

A review of audio–visual databases used for speech recognition can be found in [15]. We are using the CUAVE [14] database for all our experiments. More precisely, we use the static part of the "individuals" sequences, which consist of five repetitions of the English digits from "zero" to "nine" by each of the 36 speakers in the database, for a total of 1800 words from a ten-word vocabulary. The video is frontal, showing the upper part of the body, with little movement.

This database was chosen because it offers us a good balance between the variability in the data and the amount of time spent to run experiments, with different selection methods, numbers of features and levels of noise.

The words are spoken in sequence, with short silences in-between. Although the words are isolated, we treat the task like a continuous speech recognition problem—the recognizer receives a whole sequence comprising of five repetitions of all the digits, and has to recognize what is being said. We use a very simple syntax, which allows any combination of digits and silence, in any order.

The database was recorded in an isolated sound booth at a resolution of 720×480 with the NTSC standard of 29.97 fps [14]. The resulting video files are MPEG2 compressed. The audio is 16-bit, stereo, at a sampling rate of 44 kHz. There is also word-level labeling at millisecond accuracy, done manually, for all sequences of the database.

As the videos in the database are filmed in NTSC, there are some artifacts related to interlacing which are visible where there is movement. *Interlacing* is a technique used in video acquisition where the final frame is a blend of two *fields* acquired one after the other, each of them a complete image of the scene that is filmed, but at half the vertical resolution. Since we are particularly interested in the movement in the image, we chose to remove these artifacts through deinterlacing.

To obtain the original temporal resolution of 60 fps from the 30 fps video, we used *adaptive deinterlacing*, which is a method that detects the areas of the image where there is motion and separates the two fields only on those areas, leaving the rest of the image at full vertical resolution. This results in both high spatial and high temporal resolution. In the end, we can use the higher temporal resolution for a finer analysis of the motion of the mouth.

There are also some artifacts due to the MPEG2 compression, consisting particularly in block edges, which are sometimes visible. These kind of artifacts are impossible to eliminate, as the original detail has been lost.

B. Our AVSR System

Our speech recognition system is based on multi-stream hidden Markov models (MSHMMs), which we chose for their ability to vary the importance of each stream to the recognition.

Hidden Markov models (HMMs) [40] have the ability to model sequences of data, making them particularly well-suited for speech recognition. Multi-stream HMMs [41] have been proposed to generalize this framework for multimodal processing. They are similar to classical HMMs, with the particularity that each state contains not one, but several emission probability models, one for each stream, which are combined through a weighted product. In this way the emission likelihood b_j for state j and observation o_t at time t is expressed as [42]

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} N(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\lambda_{sjt}} \tag{10}$$

where $N(o; \mu, \Sigma)$ denotes the value in o of a multivariate Gaussian with mean μ and covariance matrix Σ. For each stream s, M_s Gaussians are used in a mixture, each weighted by c_{jsm}. In (10), the contribution of stream s is weighted by an exponent λ_{sjt} which in general can also depend on time t and state j. The latter dependency is not considered in this paper. For the experiments performed in this paper, $\lambda_{sjt} = \lambda_s$ is fixed for each signal-to-noise ratio (SNR).

Each of our word models has 32 states, tied in pairs to better model the duration of speech sound. Each Markov state is modeled with only one Gaussian with diagonal covariance. For the actual implementation we used the widely popular HTK library [42]. The word recognition rate (WRR) is computed as the number of correctly recognized words minus the number of insertions, divided by the total number of words. The alternative measure that will also be used is the word error rate $(\mathrm{WER} = 1 - \mathrm{WRR})$.

We run our experiments in a mismatched speaker-independent scenario. Training is always done only on clean conditions, while testing is done on all SNRs. We use leave-one-out cross-validation to compensate for the variability between speakers, which makes speaker-independent tests difficult. For each experiment there are 36 runs, with one speaker left aside for testing and 35 speakers used for training. At the end, we report the average of the 36 runs.

Training of the MSHMMs is done separately for each stream. The models are then joined, with transition probabilities chosen as a weighted sum of transitions from each stream, with weights being the same as the ones used in testing. At test time, the likelihood is computed as follows:

$$\log b_j(o_t^{av}) = \lambda_a \cdot \log b_j(o_t^a) + \lambda_v \cdot \log b_j(o_t^v). \tag{11}$$

Testing is done with weights λ constant in time, or fixed, with $\lambda_a + \lambda_v = 1$. We run experiments with several pairs of weights, from $(\lambda_a; \lambda_v) = (0.0; 1.0)$ to $(\lambda_a; \lambda_v) = (1.0; 0.0)$ with step 0.05, and then choose the best weights for the particular conditions of the test. Although this cannot be done in a practical setup, it gives us a good estimate of multimodal performance in practice, and our focus here is the quality of the features, and not the multimodal fusion.

C. Audio and Visual Features

Our audio features are 13 mel-frequency cepstral coefficients (MFCCs) [43], [44], with first and second temporal derivatives. Cepstral mean normalization (CMN) [45] was also applied on the audio features, to reduce the influence of the microphone and transmission system on the data, making the recognition system more robust. This is a common setup for all audio-only speech recognizers, and our focus was more on improving the quality of the visual features.

We run our experiments with additive noise in the audio, at various SNRs, from 25 dB to -10 dB. The noise that we use, babble noise, consists of continuous speech added over the clean signal. This is a particularly difficult case, as the noise has the same characteristics of the original signal.

In order to extract visual features, the region of interest (ROI), that is, a rectangle around the mouth of the speaker, needs to be located first. The ROI is extracted based on the positions of the corners of the mouth. A rectangular area is cut around the mouth, in such a way that the mouth is centered, rotated

and scaled relative to the average mouth width over each sequence. Bilinear interpolation is used to obtain the final images, all having 128×128 (16384) grayscale pixels.

For visual features we use the 2d-DCT of the mouth region, having the same size, 128×128, as the original image. To reduce the computational cost of our selection algorithm, we will only take into consideration a subset of the 16 384 coefficients. We keep only the low-frequency coefficients which are contained in the upper-left triangle of the 2d-DCT. Previous work [32] uses the same technique, starting from the assumption that only spatial frequencies with periods close to the size of the features of the mouth are relevant, and high frequencies are only noise. However, we take into account a much higher number of coefficients, 64 compared to only 20, to ensure that all the relevant information is contained in this feature set. Doubling this number to 128 only increases the computational cost, not the accuracy, showing that 64 features are indeed sufficient for the task. The DC value is also discarded as insignificant.

Previous research [39] shows that the coefficients in the odd columns of the DCT have a much higher relevance, because of the symmetry of the mouth. Using only the odd columns is equivalent to imposing horizontal symmetry to the image. Thus, the even columns of the transform are removed as a way of imposing symmetry on the ROI. In the end, this leaves us with 64 DCT coefficients out of the original 16 384 pixels.

First and second temporal derivatives are computed on all coefficients, increasing the size of the initial feature set to 192. Also, similar to CMN in the audio, the means of the visual features are removed.

We always include monomodal results for comparison, and also results with the LDA transform applied on the 192 DCT coefficients with their temporal derivatives, since this is the most commonly used method to further reduce the dimensionality of the visual feature vector.

D. Our Feature Selection Algorithm

As shown in Section II-B, mutual information has been quite extensively used as an evaluation measure for feature quality for classification. There are two reasons for this. The first comes from the definition of MI, as it can expose a dependency between two random variables, even when that dependency is nonlinear. This makes MI a much more powerful measure of dependency than correlation. The second reason comes, as has been mentioned before, from Fano's inequality, which can be interpreted in a way that shows that the bound on the probability of classification error can be lowered by choosing features with higher MI with the class label.

A large majority of MI feature selection algorithms [12], [21]–[25] aiming to find k features from an initial set of n follow the *greedy algorithm* [18], which searches for a global optimum by making the locally optimal choice at each step. Applied to feature selection with MI, it leads to the following.

1) (Initialization) Start with a complete set F of initial features and an empty set S of selected ones.
2) (MI computation) Compute $I(f;C)$, the MI between each feature $f \in F$ and the class variable C.
3) (First choice) Choose $f = \arg\max_{f \in F} I(f;C)$, set $F \leftarrow F \setminus \{f\}$, $S \leftarrow \{f\}$
4) (Greedy selection) Repeat until $|S| = k$:
   4a) (Evaluation) Compute the evaluation measure as $I(f;C) -$ penalty for each $f \in F$ based on the previously selected features $s \in S$ and the class labels
   4b) (Choice) Choose feature f which maximizes the evaluation measure, set $F \leftarrow F \setminus \{f\}$, $S \leftarrow S \cup \{f\}$
5) (End) The set S contains the *best* k features according to the evaluation measure.

The evaluation step, 4a), is the step that changes between different MI selection algorithms. As shown in Section III, most algorithms try to maximize an approximation of the MI between the set of chosen features S and the class labels, a problem which reduces to computing the redundancy between feature f, the set S and the class labels C. This approximation of the redundancy is used as a penalty on the class MI term $I(f;C)$.

The particular algorithms that we compare are maximum MI (2), MIFS (3), and CMI (5). We follow the greedy method steps detailed above, changing the redundancy penalty for each of the three algorithms. With the notation above, for maximum MI, $\text{penalty} = 0$, for MIFS $\text{penalty} = \beta \sum_{s \in S_k} I(f;s)$, while for CMI $\text{penalty} = \max_{s \in S_k} I(f;s;C)$.

Out of these three algorithms and many others that exist, only the simplest one has been previously applied to AVSR. Previous approaches to the problem of feature selection for AVSR using MI choose the features with maximum MI with the class labels [12], [13], [38]. That is, the evaluation measure is simply $\max I(f;C)$, and no measure of redundancy is taken into account. By contrast, we show that penalizing features for their redundancy can improve the recognition accuracy.

In all our following tests, we use the greedy selection method outlined here. We start by using maximum MI as an evaluation measure, and then we propose two other measures which also include a penalty for redundancy between features. Our contribution here is twofold. First, we prove from our experiments that reducing redundancy between features is essential when building a feature set. Second, we make an extensive evaluation of our proposed feature selection methods based on MI, showing both monomodal and multimodal results, for a wide range of dimensionality values. We compare our results to a common method for dimensionality reduction in AVSR, the LDA, and also the maximum MI method.

In all our experiments, MI values are approximated through probability density estimation with histograms [46]. Only two-dimensional (feature value and class variable) and three-dimensional (two feature values and class variable) probability density functions (pdfs) are estimated this way, as we consider that, with the limited number of samples available, higher-dimensional pdfs are impossible to estimate reliably. The classes that we use are groups of HMM states, which we consider to be close to the true speech units. The number of bins k to be used can be estimated either with Sturges' rule [47], [48] $k_S = 1 + \log_2 n$, as a function of the number of training samples n, or with Doane's rule [48], [49] $k_D = k_S + \log_2(1 + \gamma\sqrt{n/6})$, including the skewness of the data γ. The ideal number of bins obtained with Sturges' rule is $k_S = 18$, while with Doane's we get $k_D = 22$ for the audio, respectively $k_D = 23$ for the video. Since in our experiments there is little variation on the MI values with the number of bins, we used 20 bins for all histograms. Such

histograms are used to estimate marginal probabilities for each feature, $p(Y_i)$, joint probabilities $p(Y_i, Y_j)$, $p(Y_i, C)$ and also the three-variable joint probability $p(Y_i, Y_j, C)$. These, together with the marginal P(C), are the building blocks used to compute all MI values [20]:

$$I(Y_i; C) = \sum_{y \in Y_i} \sum_{c \in C} p(y, c) \log \frac{p(y, c)}{p(y)p(c)} \tag{12}$$

$$I(Y_i; Y_j) = \sum_{y \in Y_i} \sum_{z \in Y_j} p(y, z) \log \frac{p(y, z)}{p(y)p(z)} \tag{13}$$

$$I(Y_i; Y_j; C) = \sum_{y \in Y_i} \sum_{z \in Y_j} \sum_{c \in C} p(y, z, c) \times \log \frac{p(y, z)p(y, c)p(z, c)}{p(y)p(z)p(c)p(y, z, c)} \tag{14}$$

where y and z are histogram bins for features Y_i and Y_j, while c are the individual classes. The probabilities are computed as the proportion of occurrences in each bin relative to the total number of occurrences. For example, $p(y, c)$ is the number of occurrences of feature Y_i having a value in bin y while the sample has class c, divided by the total number of samples.

However, it should be noted that, since we are always approximating MI between high-dimensional pdfs with MI between three-dimensional pdfs, it is also possible to overestimate the redundancy between them. This means that there will be cases when features are penalized too much. This is why a parameter like β from the MIFS algorithm may be useful.

A natural question that might arise is why not use all the available features for recognition. We will also answer this question, showing that, due to the curse of dimensionality, multimodal recognition performance decreases when the dimensionality becomes too big. In the end, practice confirms the intuition that small sets of features with little redundancy perform better. Redundancy needs to be reduced, as it is possible, when not adding a penalty for it, to select features that contain the same information, to the detriment of others which would bring complementary information.

V. Results

In the following, we will first show the difference in information content between the different types of features that we use, and then we will present results with three feature selection algorithms, one based only on the maximum MI, and two others which also include a penalty for redundancy between features. We apply our selection algorithms on a set of 192 features consisting of 64 DCT coefficients, deltas and delta-deltas, as detailed in Section IV-C. From these, we select subsets going from four features all the way to the full 192 feature set, to assess how the dimensionality influences performance, and which selection algorithms performs better.

A. Mutual Information as a Measure of Feature Quality

The only algorithm that has been used before for visual feature selection in AVSR is the maximum mutual information. Basically, it means features are selected only by their MI value with the class, irrespective of previously selected features, as in (2).

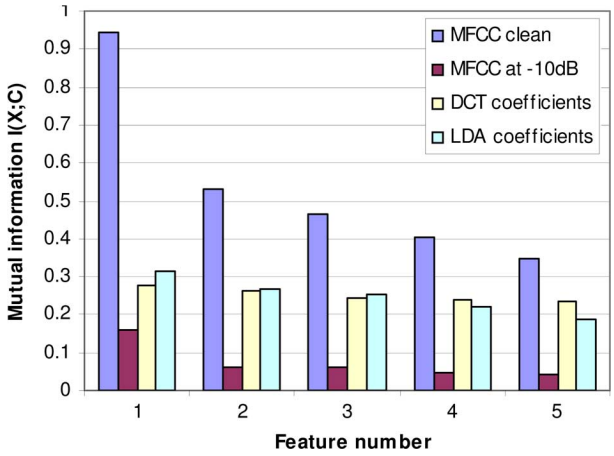

Fig. 1. The amount of relevant information (MI between the feature and the class label) for the best five features, for four types of features: clean audio, noisy audio, DCT, and LDA.

The selection algorithm is reduced to sorting the features in F by their MI value and picking only the top part of the list for inclusion in the subset S.

Fig. 1 shows the first five MI values for the best features. We analyzed four types of features: MFCCs from clean audio, MFCCs from noisy audio, DCT features and finally LDA features. What can be seen from the graph is that, as expected, the clean audio contains the highest amount of information about classes. However, when the audio is corrupted by noise, the amount of information decreases drastically, and below the level in any of the visual features. Between these visual features, the LDA coefficients have the highest MI for the first few features. However, the MI for the latter ones decreases faster in the case of LDA coefficients than in the case of the DCT.

This hierarchy is also reflected in the monomodal classification results, confirming that MI is a good measure for feature relevance.

Note that, according to the data processing inequality, the information contained in the set of LDA features should be less, or, in the best case, the same, as the information from the set of DCT coefficients. Indeed, the LDA coefficients are obtained directly through the application of a transform (the LDA) on the DCT coefficients, the same which are shown on the graph. The data processing inequality [20] claims that information cannot be created when a transform is applied on the data. So, although some of the LDA coefficients seem to contain more information than the DCT ones, on the whole, the set of LDA features will have the same information, or less.

B. Results With Maximum Mutual Information

Fig. 3(a) shows the visual-only recognition results for all MI feature selection algorithms, for feature dimensionality ranging from 4 to 192. Results for LDA are also included for comparison. Looking only at the maximum MI results, we can see that they outperform the LDA at higher dimensionality, but the LDA is better between 4 and 20 features. The maximum MI method obtains a maximum performance for around 60 features.

Fig. 3(b) shows the audio–visual recognition performance for the same features, for an audio SNR of -10 dB with babble noise. Although the audio-only accuracy in this case is only

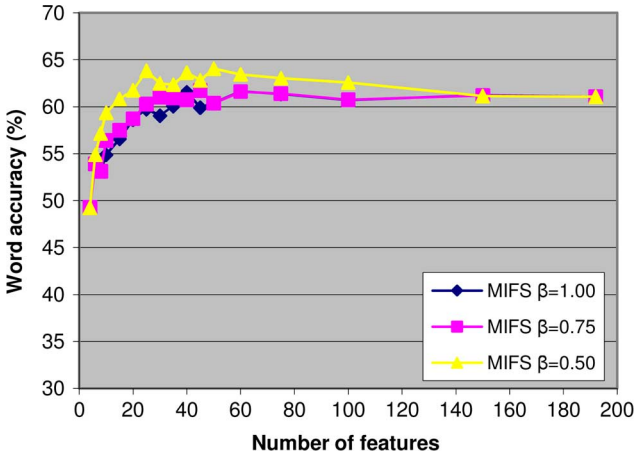

Fig. 2. Visual-only results with MIFS, using three different values for the parameter β, 0.5, 0.75, and 1.0.

22.3%, audio–visual recognition rates are in all cases better than both audio-only and visual-only ones.

There is a larger gain from multimodality at lower dimensionality values compared to higher ones. For example, for ten visual features, 5.2% are gained, while for 100, the gain is of only 1.6%. This could mean that the curse of dimensionality is impeding any gains in performance with a high number of features.

C. MMI With Weighted Redundancy Penalty

The first algorithm that we propose for selecting visual features, MIFS (mutual information feature selection) [21], penalizes features for their redundancy with other features in the selected set. The equation used to select features at each step is (3).

The parameter β is the proportion of the redundancy that is penalized by this algorithm, and is recommended by the author in [21] to be set between 0.5 and 1. The justification for this is that we are trying to penalize the redundancy of the feature Y_i with respect to the whole set S_k, that is $I(Y_i; S_k)$, which unfortunately we cannot compute. The sum $\sum_{Y_{\pi_j} \in S_k} I(Y_i; Y_{\pi_j})$ is actually an upper limit for that redundancy, reached in the case when all the features in the set S_k are disjoint, that is, MI between any of them is zero. Of course this is not the case, so the real penalty should be lower then the sum, and hence the parameter β. However, the true value of the parameter depends on the particularities of the data, and it fact, the optimal β is different at each selection step. We choose the parameter β heuristically, as in [21].

Fig. 2 shows a comparison between the results of the algorithm for visual-only recognition for three values of the parameter β, 0.5, 0.75 and 1.0. As can be seen the best is $\beta = 0.5$, which outperforms the maximum MI and the LDA features as well, as shown in Fig. 3(a).

In the audio–visual results with MIFS [Fig. 3(b)], the same tendency is seen as with maximum MI, that is, results improve more at low dimensionality than at high dimensionality when moving from single modality to multimodal processing. For example, for ten features with $\beta = 0.5$, the gain is 5.5%, while for 50 it is only 0.3%. For higher visual dimensionality values, multimodal performance is close to the visual-only, or monomodal, performance.

The results obtained with the MIFS method show how important it is to eliminate the redundancy between features. However,

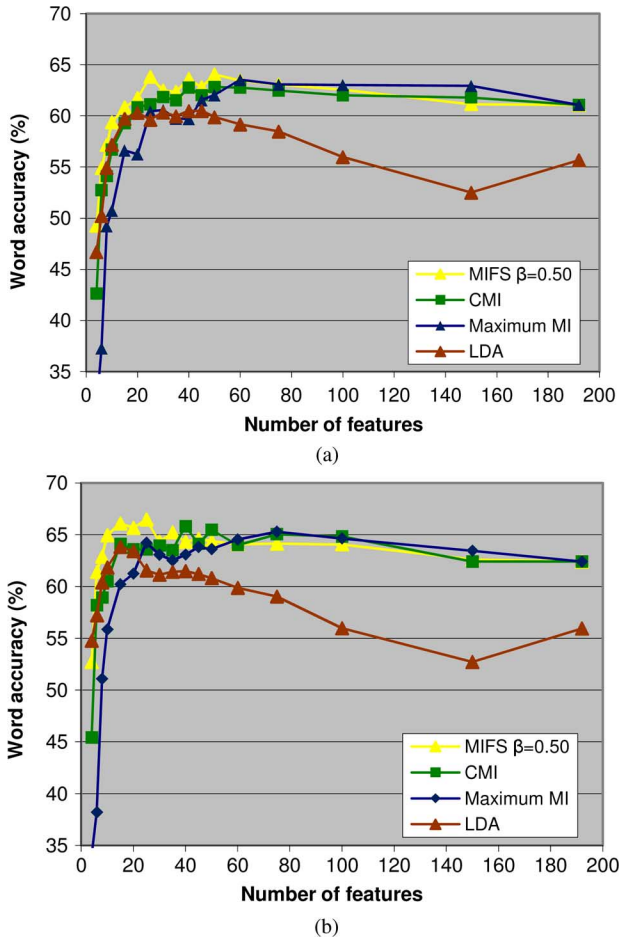

Fig. 3. Visual-only and audio–visual results with maximum MI, MIFS ($\beta = 0.5$) and CMI. The audio SNR is -10 dB with babble noise, and audio-only accuracy is only 22.3%. (a) visual-only; (b) audio–visual.

the MIFS method has an important drawback, that is, the parameter β needs to be chosen correctly for the method to give good results. In the next section we present a method which does not depend on any parameter to set the penalty for redundancy.

D. Selection With Conditional Mutual Information

Our second proposed method is to maximize the conditional mutual information (CMI), that is, the information that is added by a feature to what was already known about the class label through the other features. The equation used by the CMI algorithm is 5, which is equivalent to penalizing only *relevant* redundancy, $I(Y_i; Y_{\pi_j}; C)$. This measure can be either positive or negative [20]. When it is positive, it can be interpreted as the information about the class label that is shared, or redundant, between two features. However, when this measure is negative, it can have a different interpretation, as a measure of synergy, that is, how much information about the class label is added by taking two features together, compared to just taking them individually. This interpretation comes from the development of the joint mutual information for two features X, Y and the class label C:

$$\begin{aligned} I(X,Y;C) &= I(X;C) + I(Y;C\,|\,X) \\ &= I(Y;C) + I(X;C\,|\,Y) \\ &= I(X;C) + I(Y;C) - I(X;Y;C). \end{aligned} \quad (15)$$

Here the conditional mutual information $I(Y;C\,|\,X)$ (or respectively $I(X;C\,|\,Y)$) is not necessarily smaller than its unconditional counterpart $I(Y;C)$ (or $I(X;C)$), which indeed means that the three-way MI can be negative. When this is the case, the joint mutual information between the pair (X,Y) and the class label C, $I(X,Y;C)$, is higher than the MI sum $I(X;C)+I(Y;C)$, which means that there is synergy between the variables.

The CMI algorithm works as follows. First, for a certain candidate feature Y_i which was not included in the selected feature set yet, a corresponding feature Y_{π_j} is found, the one that is maximally redundant to it. This can be expressed either by minimizing $I(Y_i;C\,|\,Y_{\pi_j})$ or by maximizing $I(Y_i;Y_{\pi_j};C)$, as can be seen from (5). The second step is choosing the feature Y_i which adds the most information about C to its most redundant corresponding feature, Y_{π_j}. This basically assumes that if the feature Y_i adds a lot of information even compared to its most redundant counterpart, it will also add information not present in the whole set S_k.

The CMI $I(Y_i;C\,|\,Y_{\pi_j})$ represents the information about C that is brought by Y_i that is supplementary to what was already known through Y_{π_j}. Obviously the measure that should be used in fact is $I(Y_i;C\,|\,S_k)$, the information about C brought by Y_i that is complementary to the information in the whole set of selected features, S_k. However, because of the high dimensionality, this cannot be computed, so we have to rely on approximations.

Visual-only results with the CMI selection algorithm are included in Fig. 3(a). The performance is on the same level as the LDA for the low dimensionality values. Compared to the maximum MI algorithm performance is better, however compared to MIFS it is a little worse. The same tendencies can be seen on the audio–visual performance graph [Fig. 3(b)].

Although the CMI algorithm only chooses features which are complementary to the already selected set, it performs a little worse than MIFS. It is possible that there are cases where the CMI algorithm penalizes features too much, that is, it considers them as having too much redundancy, although by themselves they contain a lot of information.

E. Performance in Clean Conditions

The previous analysis was done only with corrupted audio, with a SNR of -10 dB. However, our goal is to have a system which performs well across all conditions. In the following, we will present results at other SNRs as well.

Fig. 4 show audio–visual performance with all selection algorithms and LDA, compared to audio-only. As can be seen, the CMI method is the best in this case, although by a very slight margin. For clean audio, the monomodal recognition accuracy is already very high, at 98.3%. Multimodal recognition improves this, irrespective of the visual features' dimensionality. All MI selection algorithms further improve performance compared to the LDA, although by very little in absolute terms. However, given the very small number of recognition errors, the relative reduction in error is impressive, 16.6% passing from LDA to MIFS for example. Although there is a lot of variability in the results, the same tendency is seen across all dimensionality values, that is, MI-based methods are better than the LDA. Accuracy to one decimal place here is warranted since a small accuracy variation can have a large influence on the number of errors.

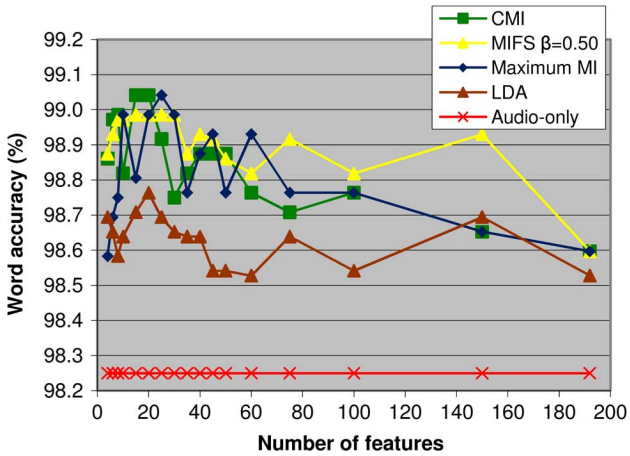

Fig. 4. Audio–visual results with clean audio, compared to audio-only.

Between the MI-based methods, it is difficult to pick a clear winner. The maximum performance is obtained with either CMI or maximum MI, depending on dimensionality. However, it is clear that the optimal dimensionality is between 15 and 25, much less than for corrupted audio.

The fact that less visual features are necessary to obtain an optimal performance for clean audio could be explained by the redundancy between the audio and visual features, which is lost when the SNR decreases. This means that, when the audio is degraded, information that was common to audio and video is now only present in the video, and more features are necessary in order to include all this information in the feature set.

F. Performance at 20, 10, and 0 dB SNR

Having analyzed the influence of the feature extraction methods at the two extreme SNRs, in this section we will also show results three in-between SNR values: 20, 10, and 0 dB. The results are presented in Fig. 5(a), (b), and (c).

At 20 dB, performance is still very high, mostly above 97% for all MI methods, which all show better performance than the LDA. Both CMI and MIFS show an improvement over maximum MI, which, although small in absolute terms (0.7%), amounts to an 8% relative reduction in the error rate. Here CMI performs best, at a dimensionality of only 6.

At a SNR of 10 dB, the audio is quite corrupted and performance decreases to 85.3% for monomodal recognition. The gain from multimodal recognition is impressive, 6.7% in absolute terms, or a relative reduction in the number of errors of 45.6%. Again the best-performing method is MIFS. Both MI-based methods are better then the LDA at all dimensionality values.

The ideal number of visual features, the dimensionality that gives the highest performance for AV recognition is here only 8. This is true for the MIFS features, which, when used alone for monomodal recognition, give the best results when the dimensionality is 25. The reduction in the number of features which give the best result compared to video-only may be explained by the curse of dimensionality, but also, as mentioned before, by the fact that there is redundancy with the audio, which means

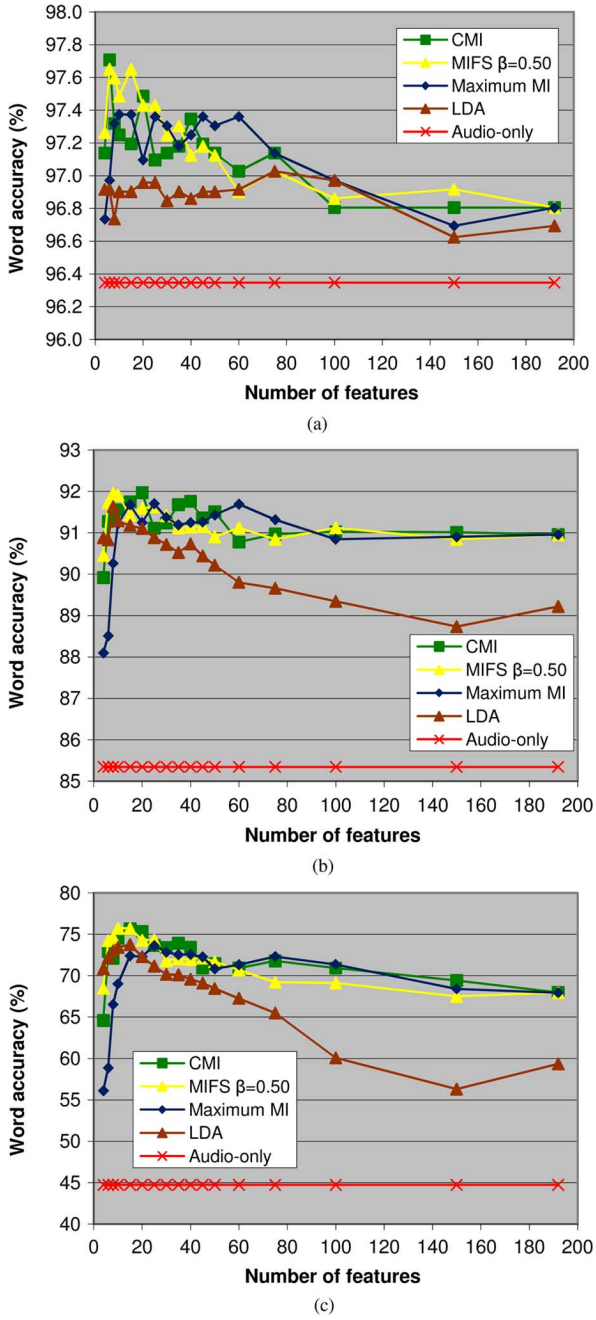

Fig. 5. Audio–visual results with three audio SNRs with babble noise, compared to audio-only. (a) 20 dB; (b) 10 dB; (c) 0 dB.

that only a reduced number of features bringing complementary information is actually required to augment the audio.

Looking at the performance of the maximum MI algorithm, we can see that here it is slightly lower overall than that of CMI of MIFS. Although maximum MI has a higher performance for 60 and 80 features, this is irrelevant since the optimal performance for this SNR is obtained for 15 to 25 visual features.

For 0 dB, performance decreases even more. Comparing the MI selection methods, we see that overall the gap between CMI and MIFS compared to maximum MI increases. There is a difference of 3% between the best performance obtained with CMI and MIFS compared to maximum MI. As was the case before, maximum MI slightly outperforms the other two at higher dimensionality, but at a lower accuracy than with small dimensionality.

Overall, we see that the ideal number of visual features is between 10 and 20, lower than the typical average number of 40 used in previous work. Our results show that with a good selection algorithm it is possible to obtain better results with less visual features, across a wide range of SNRs. For example, the improvement in going from 40 to 15 features for MIFS is an average of 9% relative reduction in WER. Having a lower number of features has the additional advantage of requiring less storage and computing resources, leading to a more efficient application. Unfortunately, the ideal number of features depends on the SNR, so it should be chosen on a held-out set recorded if possible in conditions similar to those in which the application will be deployed, that is, similar to those of the final test set.

The methods that we propose lead to an increase in performance and a decrease in the required number of features compared to LDA and maximum MI. This is because by including the redundancy of the features as a penalty in the quality measure we obtain feature sets which are more compact and at the same time more relevant to our classification problem. The fact that the performance decreases for higher dimensionality with all methods can be attributed to the curse of dimensionality, and more precisely the fact that the number of samples available for training is limited.

G. Performance Across All SNRs

Overall, we see that the two methods which penalize features proportionally to their absolute redundancy perform better than the others, across a wide range of SNRs. We will present now results for all SNRs, from clean to -10 dB, for babble noise.

As seen in the previous subsections, results vary quite a lot with the dimensionality of the video features. The dimensionality that gives the best performance for clean audio is not the same as the dimensionality that is optimal for noisy audio. To have an even field for comparison, we chose a dimensionality of 15 for all the types of visual features we use. Fig. 6(a) shows the audio–visual performance at all SNRs. Because the variations in absolute accuracy values are very small for higher SNRs, it would seem that all algorithms perform more or less the same. However, when translated in relative error reduction terms, the differences are significant.

In Fig. 6(b) we present the percentage reduction in WER of each algorithm relative to the audio WER. This graph shows much more clearly the gains of CMI and MIFS over LDA and maximum MI, for all SNRs. Averaging over all SNRs, MIFS has a 5.5% advantage in WER reduction over maximum MI, while CMI has 4.6%. MIFS is best in very noisy conditions, from -10 dB up to 0 dB, while CMI features perform better when paired with cleaner audio.

In the end, MIFS and CMI are clearly performing better than the visual feature selection methods typically used in AVSR, LDA and maximum MI, achieving significant reductions in WER for almost all SNRs. The reason for their performance is the fact that they lead to more compact feature sets, with less redundancy between features. Still, some information will be lost through this process, and this lost information might explain the differences between CMI and MIFS. If

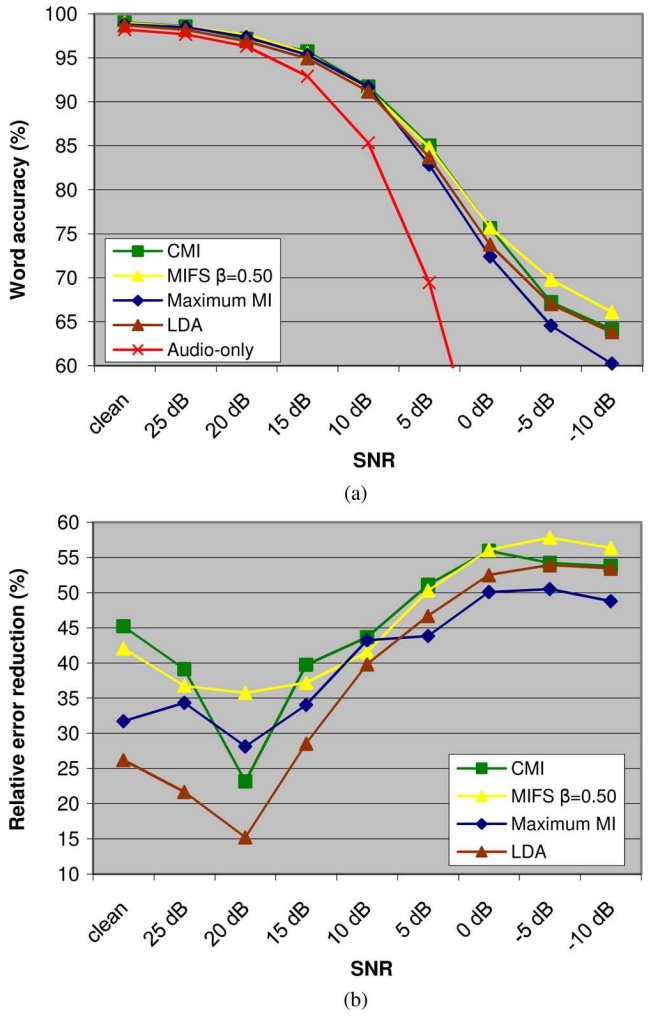

Fig. 6. Audio-visual results at all SNRs, with babble noise. The dimensionality of all visual features is 15. (a) Word accuracy by SNR, (b) Relative error reduction (%).

some information complementary to the audio is kept to the detriment of information redundant with respect to the audio, this would lead to better performance in clean conditions, and worse performance with noise. This could be the reason why CMI features perform better in clean conditions, while MIFS features are better with noisy audio.

VI. Conclusion

Throughout this paper we strived to validate our results through leave-one-out testing, by using varying noise levels and also various selection methods and dimensionality values for the features.

Our contribution here is two-fold. First, we propose two methods of visual feature selection for AVSR, methods based on maximizing mutual information with the classes while at the same time minimizing redundancy. We proved that penalizing features for their redundancy is important and obtained significant performance gains compared to the state of the art.

Second, we perform an extensive analysis of information theoretic feature selection methods applied on visual features for AVSR, for different audio SNRs, with realistic babble noise. As opposed to many approaches in the literature, where typically a visual feature vector of dimensionality 40 is used, we present results across a wide range of dimensionality values. We prove that, in many cases, a small feature vector can outperform higher-dimensional ones, proving that obtaining low-dimensional features is always desirable.

Our work shows that, for audio–visual speech recognition, mutual information is not only a good measure for the relevance of features, but also a good way of estimating the redundancy between them. Penalizing features for their redundancy leads to better sets of features, outperforming even the LDA. Of the two proposed methods, CMI works a little better with clean audio, while MIFS performs better in noisy conditions.

Knowing which features are relevant and which are not, which features contain the same information and which complement each other can be very useful not only for building better feature sets, but also for our own understanding of the problem. We could, for example, globally evaluate the quality of feature extraction methods, and choose the one which gives us features having the most information. The information theoretic analysis might also show that features obtained with two different transforms are complementary, and using more than one feature type might increase performance. It is common for example in AVSR to use both appearance-based features and shape features together, as they complement each other well. MI is a tool that could be used to identify similar situations where more than one feature type may be useful.

References

[1] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, 1998.
[2] B. Zitova and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, pp. 977–1000, 2003.
[3] J. Richards and X. Jia, *Remote Sensing Digital Image Analysis*. New York: Springer, 2006.
[4] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. Cambridge, MA: MIT Press, 2004.
[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
[6] N. Fox, R. Gross, J. Cohn, and R. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 701–714, 2007.
[7] R. E. Bellman, *Adaptive Control Processes: A Guided Tour.*. Princeton, NJ: Princeton Univ. Press, 1961.
[8] J. Hershey and J. R. Movellan, "Audio vision: Using audio–visual synchrony to locate sounds," *Neural Inf. Process. Syst.*, pp. 813–819, 1999.
[9] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," *Neural Inf. Process. Syst.*, pp. 814–820, 2000.
[10] H. J. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," presented at the ACM Multimedia, Juan-les-Pins, France , Dec. 1–6, 2002.
[11] J. Fisher, III, T. Darrell, W. Freeman, and P. Viola, "Learning joint statistical models for audio–visual fusion and segregation," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 772–778, 2000.
[12] P. Scanlon, G. Potamianos, V. Libal, and S. M. Chu, "Mutual information based visual feature selection for lipreading," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Jeju Island, Korea, Oct. 4–8, 2004, pp. 2037–2040.
[13] I. Arsic and J. P. Thiran, "Mutual information eigenlips for audio–visual speech recognition," in *Proc. 14th Eur. Signal Processing Conf. (EUSIPCO)*, 2006.
[14] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 11, pp. 1189–1201, 2002.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio–visual speech," *Proc. IEEE*, vol. 91, no. 9, 2003.
[16] T. Drugman, M. Gurban, and J. Thiran, "Relevant feature selection for audio–visual speech recognition," presented at the 9th Int. Workshop Multimedia Signal Processing (MMSP), Chania, Crete, Greece, Oct. 1–3, 2007.
[17] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA: Kluwer Academic, 1998.
[18] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*. New York: McGraw-Hill, 2003.
[19] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 765–773, 2004.
[20] T. Cover and J. Thomas, *Elements of Information Theory.*, ser. Wiley Series in Telecommunications. New York: Wiley, 1991.
[21] R. Battiti, "Using mutual information for selecting features in supervised neural net working," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, 1994.
[22] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.
[23] N. Kwak and C. H. Choi, "Improved mutual information feature selector for neural networks in supervised learning," in *Proc. IEEE Int. Joint Conf. Neural Networks*, 1999, vol. 2, pp. 1313–1318.
[24] M. Tesmer and P. Estevez, "AMIFS: Adaptive feature selection by using mutual information," in *Proc. IEEE Int. Joint Conf. Neural Networks*, 2004, vol. 1, p. 308.
[25] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 281–288.
[26] R. Reilly and P. Scanlon, "Feature analysis for automatic speechreading," in *Proc. Workshop Multimedia Signal Processing*, 2001, pp. 625–630.
[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and minredundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, 2005.
[28] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[29] R. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.
[30] C. Rao, "The utilization of multiple measurements in problems of biological classification," *J. Roy. Statist. Soc.*, ser. B, vol. 10, pp. 159–203, 1948.
[31] J. Luettin and N. A. Thacker, "Speechreading using probabilistic models," *Comput. Vis. Image Understand.*, vol. 65, no. 2, pp. 163–178, 1997.
[32] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. Int. Conf. Image Processing*, 1998, vol. 3, pp. 173–177.
[33] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio–visual speech recognition," in *Proc. Int. Conf. Spoken Language Processing*, 2000, vol. III, pp. 20–23.
[34] H. H. Yang, S. V. Vuuren, S. Sharma, and H. Hermansky, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Commun.*, vol. 31, no. 1, pp. 35–50, 2000.
[35] P. Scanlon, D. Ellis, and R. Reilly, "Using mutual information to design class-specific phone recognizers," presented at the Eurospeech, Geneva, Switzerland, 2003.
[36] P. Scanlon, D. Ellis, and R. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 803–812, 2007.
[37] D. Ellis and J. Bilmes, "Using mutual information to design feature combinations," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2000, vol. 3, pp. 79–82.
[38] P. Scanlon, *Audio and Visual Feature Analysis For Speech Recognition*. Dublin, Ireland: University College, 2005.
[39] G. Potamianos and P. Scanlon, "Exploiting lower face symmetry in appearance-based automatic speechreading," presented at the Int. Conf. Audio–Visual Speech Processing, British Columbia, Canada, Jul. 24–27, 2005.
[40] L. R. Rabiner, "A tutorial on Hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, 1989.
[41] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 426–429.
[42] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, U.K.: Entropic, Ltd., 1999.
[43] P. Mermelstein, , R. Chen, Ed., "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recogn. Artific. Intell.*. New York: Academic Press, 1976.
[44] S. Davis and P. Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 357–366, 1980.
[45] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.
[46] C. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press, 2005.
[47] H. Sturges, "The choice of a class-interval," *J. Amer. Statist. Assoc.*, vol. 21, pp. 65–66, 1926.
[48] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley, 1992.
[49] D. Doane, "Aesthetic frequency classification," *Amer. Statist.*, vol. 30, pp. 181–183, 1976.

Mihai Gurban (M'09) was born in Timisoara, Romania, in 1979. He received the Computer Science engineer diploma from the Politehnica University in Timisoara in 2003 and the Ph.D. degree from the Swiss Federal Institute of Technology (EPFL), in Lausanne, Switzerland, in 2009.

He continues work at EPFL as a Postdoctoral Research Fellow at the Signal Processing Laboratory (LTS). His scientific interests include multimodal signal processing, dimensionality reduction, image processing, speech recognition, and machine learning.

Jean-Philippe Thiran (SM'03) was born in Namur, Belgium, in 1970. He received the Elect.Eng. and Ph.D. degrees from the Universite Catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in 1993 and 1997, respectively.

He joined the Signal Processing Laboratory (LTS) of the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in February 1998 as a Senior Lecturer. Since January 2004, he has been an Assistant Professor, responsible for the Image Analysis Group. His current scientific interests include image segmentation, prior knowledge integration in image analysis, partial differential equations and variational methods in image analysis, multimodal signal processing, medical image analysis, including multimodal image registration, segmentation, computer-assisted surgery, and diffusion MRI. He is author or coauthor of five book chapters, 73 journal papers, and some 130 peer-reviewed papers published in proceedings of international conferences. He holds four international patents.

Dr. Thiran was Co-Editor-in-Chief of *Signal Processing* (Elsevier Science) from 2001 to 2005. He is currently an Associate Editor of the *International Journal of Image and Video Processing* (Hindawi), and member of the Editorial Board of *Signal, Image and Video Processing* (Springer). He was the General Chairman of the 2008 European Signal Processing Conference (EUSIPCO 2008). He is a member of the MLSP and IVMSP technical committees of the IEEE Signal Processing Society.