Relevant Feature Selection for Audio-Visual Speech Recognition

Thomas Drugman
Faculté Polytechnique de Mons
Mons, Belgium
Email: thomasdrugman@hotmail.com

Mihai Gurban and Jean-Philippe Thiran
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
Email: {Mihai.Gurban, JP.Thiran}@epfl.ch

***Abstract*— We present a feature selection method based on information theoretic measures, targeted at multimodal signal processing, showing how we can quantitatively assess the relevance of features from different modalities. We are able to find the features with the highest amount of information relevant for the recognition task, and at the same having minimal redundancy. Our application is audio-visual speech recognition, and in particular selecting relevant visual features. Experimental results show that our method outperforms other feature selection algorithms from the literature by improving recognition accuracy even with a significantly reduced number of features.**

I. Introduction

The ability to rank features by their relevance and select only the best ones is very useful in pattern recognition. Here, by "relevance" we mean a feature's importance for a particular classification task. As the dimensionality of the feature vector increases, it becomes more and more difficult to accurately model the data, as an ever increasing number of samples is required. This is referred to in the literature as the "curse of dimensionality". Selecting only the most relevant features, as a means of dimensionality reduction, can increase the performance of a classifier, reduce training and testing times, as well as computational and memory requirements, and improve our understanding of the data.

In the case of multimodal signal processing, and in particular audio-visual speech recognition (AVSR), the problem of feature selection becomes even more important. AVSR [1] uses images of the speaker's mouth in order to decrease error rates when the audio is noisy. While the types of audio features for speech recognition are more or less established, in the visual domain there still is a lot to be gained through feature selection. Although they contain less information than the audio features, visual ones typically have higher dimensionality. The dynamic nature of speech requires, as for the audio, the inclusion of temporal information lying in the variation between frames, which further increases the dimensionality of the visual feature vector.

The information theoretic measures that we use can not only assess the relevance of one individual feature, but also approximate the information added by it with respect to a whole set of other features, by computing the redundancy between them. Our two selection methods differ from previous work in the way that the redundancy between features is computed. We also present interesting results about the individual relevance of each feature type, and how it is tied with recognition accuracy. Finally, to the best of our knowledge, these feature selection methods had not been applied on AVSR before. Our results confirm their usefulness for AVSR, and in fact, for any multimodal classification task.

Our paper is structured as follows. In Section II we give an overview of feature selection in the context of classification. Section III details our method and its novelty w.r.t. previous work. In Section IV we present our results, showing a quantitative comparison of the relevance of different types of features and also the performance obtained with several feature selection algorithms. Section V concludes our paper.

II. Background

A. Feature selection

In this section we present previous work on feature selection algorithms [2]. The aim is to choose from a set F of n features, a subset S of m features, such that S retains most of the information in F relevant for the classification task. Since the number of possible subsets, $\binom{n}{m}$, is usually too high to allow the processing of each candidate, this leads to iterative "greedy" algorithms which choose features one by one according to some measure. In this respect, our focus is on information theoretic measures, as they can directly assess the relevance of a feature.

The mutual information (MI) $I(Y;C)$ between a feature Y and the class labels C is such a measure, extensively used for feature selection in classification. It represents the reduction in the uncertainty of C when Y is known. A high MI means that the feature is relevant for our particular classification problem.

Computing MI from data requires the estimation of probability densities, which can not be accurately done in high dimensions. This is why a majority of feature selection algorithms use measures based on up to three variables (two features plus the class label).

We will now briefly present a few information-theoretic feature selection methods. Let $F = \{Y_1, Y_2 \dots Y_n\}$ be the initial set of features. Let $\{\pi_1, \pi_2 \dots \pi_m\}$ be a permutation on a subset of dimension m of the set of feature indices $\{1 \dots n\}$. Then the set of selected features can be written as $S = \{Y_{\pi_1}, Y_{\pi_2} \dots Y_{\pi_m}\} \subset F$.

The simplest information-theoretic criterion to select a feature at step $k+1$ is [2], [3]:

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

where $S_k = S_{k-1} \cup \{Y_{\pi_k}\}$ is the set of features selected at step k.

However, this method ranks each feature's relevance individually, irrespective of previous choices. In order to have a maximum of information with a small number of features, any redundancy should be eliminated.

A possible way of doing this is to penalize a feature's importance by a proportion of its summed redundancy with the already chosen features (the MIFS algorithm [4]):

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} \left[I(Y_i; C) - \beta \sum_{Y_{\pi_j} \in S_k} I(Y_i; Y_{\pi_j}) \right] \quad (2)$$

A similar approach is to penalize the average redundancy [5]:

$$Y_{\pi_{k+1}} = \arg\max_{Y_i \in F \setminus S_k} \left[I(Y_i; C) - \frac{1}{|S_k|} \sum_{Y_{\pi_j} \in S_k} I(Y_i; Y_{\pi_j}) \right] \quad (3)$$

where $|S_k|$ is the size of set S_k.

Another family of information theoretic feature selection algorithms uses the conditional mutual information (CMI) as a measure [6], $I(X;C|Y) = I(X,Y;C) - I(Y;C)$. 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:

$$\begin{aligned} Y_{\pi_{k+1}} &= \arg\max_{Y_i \in F \setminus S_k} \left[\min_{Y_{\pi_j} \in S_k} I(Y_i; C | Y_{\pi_j}) \right] \quad (4) \\ &= \arg\max_{Y_i \in F \setminus S_k} \left[I(Y_i; C) - \max_{Y_{\pi_j} \in S_k} I(Y_i; Y_{\pi_j}; C) \right] \end{aligned}$$

using $I(X;Y;C) = I(Y;C) - I(Y;C|X)$ [7]. For a certain Y_i, the particular Y_{π_j} is found with which Y_i is most redundant. By taking the maximum over this CMI, the feature that adds the most relevant information to the set S_k is found.

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 [7]):

$$\begin{aligned} I(S;C) &= I(Y_{\pi_1}, Y_{\pi_2}, \ldots, Y_{\pi_m}; C) \\ &= \sum_{j=1}^{m} I(Y_{\pi_j}; C | Y_{\pi_1}, \ldots, Y_{\pi_{j-1}}) \quad (5) \\ &= \sum_{j=1}^{m} \left[I(Y_{\pi_j}; C) - I(Y_{\pi_j}; C; Y_{\pi_1}, \ldots, Y_{\pi_{j-1}}) \right] \end{aligned}$$

Since not all subsets can be tested, an iterative algorithm could maximize the terms of this sum one by one. Since Y_{π_j} is the particular Y_i that maximizes the j^{th} term of the sum, all previously mentioned criteria (Eq. 2, 3, 4) 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 $I(Y_i; C; Y_{\pi_1}, \ldots, Y_{\pi_{j-1}})$ between Y_i, S_{j-1} and the class labels C.

B. Feature selection with MI in speech recognition

Mutual information has been used before in speech recognition as a criterion for feature selection. In [8], conditional mutual information is used to find combinations of audio feature streams. In [9], mutual information is used to select visual features for AVSR from a set of DCT coefficients. Two criteria are used - either the maximum mutual information as shown in Eq. 1, or the joint mutual information $I(Y_i, Y_j; C)$.

Our method uses two criteria, the one in Eq. 4, and a new criterion based on clustering redundant features. We will present them in detail in the next section and then, in the results section, we will show the performance improvements that they bring over similar methods from the literature.

III. Our proposed method

This section begins by introducing the way we estimate temporal derivatives of visual features. We then present our multistream classifier dedicated to AVSR. Finally, we describe the way we compute MI and our feature selection methods.

A. Temporal derivatives

After feature extraction, speech recognition systems generally take speech dynamics into account by incorporating some temporal information into the feature vector. Typically, the difference between current and previous frame values (or the k^{th} previous frame) is computed. Our temporal derivative is a weighted sum of the N previous values, where weights decrease exponentially with time as follows:

$$\Delta_i = x_i - \frac{1-\alpha}{\alpha(1-\alpha^N)} \sum_{j=1}^{N} \alpha^j x_{i-j} \quad (6)$$

This approach is similar to a first order filter whose time constant is regulated by α. Our experiments have shown that α=0.85 leads to the best results, which corresponds to a time constant close to 120ms.

Our experiments show that this method has a significant advantage in terms of performance when compared to typical first and second temporal derivatives.

B. Multistream classifier for AVSR

Our classifiers are multi-stream Hidden Markov Models (HMMs). Training is done independently for audio and video, and the model parameters are then combined to obtain a two-stream recognizer. The emission likelihood b_j for state j and observation o_t at time t is the product of likelihoods from each modality s weighted by stream exponents λ_s [10]:

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} \mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\lambda_s} \quad (7)$$

where $N(o;\mu,\Sigma)$ is the value in o of a multivariate gaussian with mean μ and covariance matrix Σ. M_s gaussians are used in a mixture, each weighed by c_{jsm}. The product in eq. 7 is in fact equivalent to a weighted sum in logarithmic domain. In practice, the weights λ_s should be tied to stream reliability, such that, when environment conditions (e.g. SNR) change, they can be adjusted to emphasize the most reliable modality. This dynamic adjustment can be based on the dispersion of the state emission posterior probabilities for each stream, as computed through Bayes's rule from the likelihoods. For the moment, however, λ_s is chosen manually.

C. MI Computation

To estimate the MI values, we opted for a histogram approach. We discretized the probability density function of each feature by finding its extreme values over the whole database and partitioning the interval into bins. Two and three-dimensional histograms (two features and the class label) were also computed. The number of bins was here empirically chosen as a trade-off between an adequately high number of bins for accurate estimation, and sufficient samples per bin. The class labels that we use for computing MI correspond to groups of HMM states representing speech phonemes.

D. Feature selection techniques

We propose two methods of feature selection. Both will maximize an individual feature's mutual information with the class labels, while at the same time removing redundancy between selected features. This is achieved with a greedy algorithm, picking a new feature at each step, as detailed in section II-A. The particularity of our methods is the fact that only the "relevant" redundancy, that is tied to the class labels, is taken into account.

Our first method uses Eq. 4. The most informative feature is chosen each time, provided that it also has little redundancy with the other chosen features.

Our second method, Selection by Redundant Features Clustering (SRFC), improves on this idea, reducing the redundancy even more. The algorithm is as follows. First, a feature is selected according to Eq. 4. Then, it is assigned to the same group as $Y_{\pi_j} = \arg\max I(Y_i; Y_{\pi_j}; C)$ if the MI is positive. Here, a positive sign means there is some redundancy, while a negative one means there is none. If the max is negative, a new group is created containing just Y_i. However, in the case of positive MI, we can refuse the selection of a new feature if its assigned group is "full", that is, there is too much redundancy inside it. This exclusion criterion is:

$$\sum_{Y_{\pi_j} \in Group} I(Y_i; Y_{\pi_j}; C) > I(Y_i; C) \tag{8}$$

In this way, if a feature satisfies the exclusion criterion, it is assumed that most of the information it conveys is already present in the group, distributed among its members. While our first technique uses a single feature for estimating redundancy, this method takes into account the whole selected subset redundancy by clustering it into internally redundant groups.

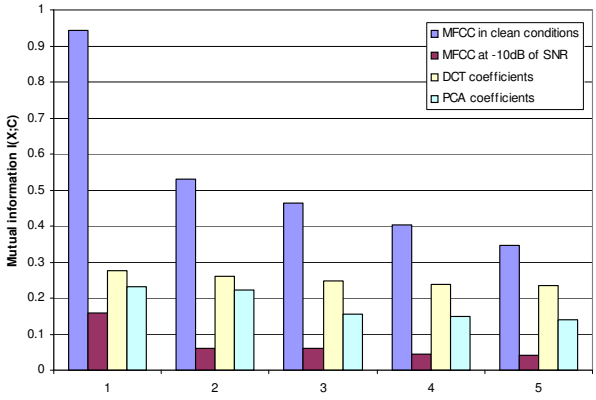

Fig. 1. Comparison of mutual information I(X;C) between different multimodal features

IV. Experimental results and discussion

A. Implementation details

For our experiments, we use sequences from the CUAVE audio-visual database [11]. It consists of 36 speakers uttering digit sequences. We use only the static part of the database, that is, 5 repetitions of the 10 digits. The region of interest (ROI) that we use is the mouth of the speaker, scaled and rotated, so that all the mouths have approximately the same size and position.

We use the HTK library [10] for the HMM implementation. Our word models have 8 states with one diagonal-covariance gaussian per state. The silence model has 3 states with 3 gaussians per state. Two streams are used, audio and video. The grammar consists of any combination of digits with silence in-between. The accuracy that we report is the number of correctly recognized words minus insertions, divided by the total number of test words.

The features that we extract from the audio stream are 13 Mel Frequency Cepstral Coefficients (MFCCs), together with their first and second temporal derivatives. As video features, we tested both Discrete Cosine Transform (DCT) and Principal Components Analysis (PCA) features computed on the ROI, with temporal derivatives as outlined in section III-A.

The testing method is as follows. The database is split into a training set of 30 sequences, and a testing set of 6. After recognition is performed, the process is repeated five times with different training and testing sets. In this way, all individual speakers are used once for testing, and five times for training. The 6 results are averaged at the end.

B. Feature quality for AVSR

We now present a comparison in terms of relevant information brought by different features commonly used in AVSR. Fig. 1 shows each individual feature's mutual information with the class label for the five best features of four categories. The two first categories are MFCCs extracted on the audio

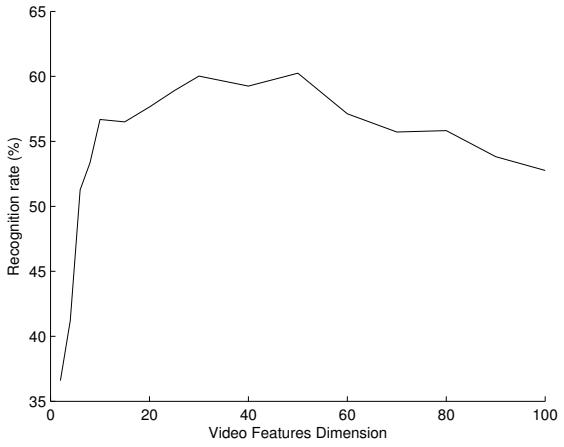

Fig. 2. Effect of dimensionality on a classifier performance for a multimodal application

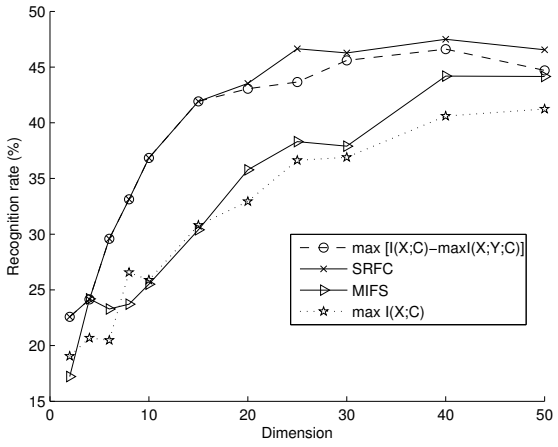

Fig. 3. Performance in video-only recognition according to dimensionality for different feature selection methods

signal for two noise levels: clean conditions and -10dB of SNR (with white Gaussian noise added). A strong decrease of the audio features quality can be noted, justifying the need to complement the corrupted audio with visual information. The last categories are two of the commonly used visual features for AVSR: DCT and PCA, with DCT clearly outperforming PCA. However, note that Fig. 1 does not show the amount of redundancy among features inside a category, but just emphasizes the intrinsic relevant information of each feature.

C. The need for feature selection in AVSR

As discussed earlier, classifier performance degrades when dimensionality becomes too high. This is even more true in the case of multimodal applications and AVSR in particular, since each modality increases the dimensionality that needs to be handled. Fig. 2 shows the audio-visual word recognition rate with corrupted audio (-10db SNR) versus the visual features dimensionality for our AVSR system. Performance peaks and then decreases as the number of features is increased.

D. The performance of our feature selection method

We compare here our feature selection method described in section III with two other methods from the literature, also described in section II-A (Eq. 1 and 2). Results are shown in Fig. 3 for video-only speech recognition. Between our two proposed methods, SRFC is better performing by a small margin. The two methods from the literature perform worse, especially at lower dimensionality. The results clearly prove that computing relevant redundancy, as opposed to just redundancy, leads to better features.

V. Conclusions and future work

We have shown that reducing the redundancy between features when selecting the most relevant ones is important for audio-visual speech. Our innovative feature selection algorithms lead to an improved performance with a small number of selected features.

As future work, we plan to improve our feature selection method, focusing on removing redundancy not only from visual features, but also from the audio and between the two modalities. We also plan to allow stream weights to vary dynamically, according to each stream's reliability.

Acknowledgement

This work is supported by the Swiss National Science Foundation through the IM2 NCCR.

References

[1] 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. MIT Press, 2004.

[2] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers, 1998.

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

[4] R. Battiti, "Using mutual information for selecting features in supervised neural net working," *IEEE Transactions on Neural Networks*, vol. 5(4), 1994.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(8), 2005.

[6] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5), pp. 1531–1555, 2004.

[7] T.Cover and J.Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, New York, 1991.

[8] D. Ellis and J. Bilmes, "Using mutual information to design feature combinations," in *Proceedings of ICSLP*, vol. 3, 2000, pp. 79–82.

[9] P. Scanlon, G. Potamianos, V. Libal, and S. M. Chu, "Mutual information based visual feature selection for lipreading," *ICSLP*, pp. 2037–2040, 2004.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, Entropic Ltd., 1999.

[11] 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 JASP*, vol. 2002(11), pp. 1189–1201, 2002.