# Multi-pose lipreading and Audio-Visual Speech Recognition

Virginia Estellers* and Jean-Philippe Thiran

Signal Processing Laboratory LTS5, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Email: Virginia Estellers*- virginia.estellers@epfl.ch; Jean-Philippe Thiran - jp.thiran@epfl.ch;

*Corresponding author

## Abstract

In this paper we study the adaptation of visual and audio-visual speech recognition systems to non-ideal visual conditions. We focus on overcoming the effects of a changing pose of the speaker, a problem encountered in natural situations where the speaker moves freely and does not keep a frontal pose with relation to the camera. To handle these situations, we introduce a pose normalization block in a standard system and generate virtual frontal views from non-frontal images. The proposed method is inspired by pose-invariant face recognition and relies on linear regression to find an approximate mapping between images from different poses. We integrate the proposed pose normalization block at different stages of the speech recognition system and quantify the loss of performance related to pose changes and pose normalization techniques. In audio-visual experiments we also analyse the integration of the audio and visual streams. We show that an audio-visual system should account for non-frontal poses and normalization techniques in terms of the weight assigned to the visual stream in the audio-visual classifier.

## Introduction

The performance of automatic speech recognition (ASR) systems degrades heavily in the presence of noise, compromising their use in real world scenarios. In these circumstances, ASR systems can benefit from the use of other sources of information complementary to the audio signal and yet related to speech. Visual speech constitutes such a source of information. Mimicking human lipreading, visual ASR systems are designed to recognize speech from images and videos of the speaker's mouth. This fact gives rise to audio-visual

automatic speech recognition (AV-ASR), combining the audio and visual modalities of speech to improve the performance of audio-only ASR, especially in presence of noise [1, 2]. In these situations we cannot trust the corrupted audio signal and must rely on the visual modality of speech to guide recognition. The major challenges that AV-ASR has to face are, therefore, the definition of reliable visual features for speech recognition and the integration of the audio and visual cues when taking decisions about the speech classes.

A general framework for AV-ASR [3] has been developed during the last years, but for a practical deployment the systems still lack robustness against non-ideal working conditions. Research has particularly neglected the variability of the visual modality subject to real scenarios, i.e non-uniform lighting and non-frontal poses caused by natural movements of the speaker. The first studies on genuine AV-ASR applications with realistic working conditions [4, 5] applied directly the systems developed for ideal visual conditions, obtaining poor performances and failing to exploit the visual modality in the multi-modal system. These works pointed out the necessity of new visual feature extraction methods robust to illumination and pose changes.

In lipreading systems, the variations of the mouth's appearance caused by different poses are more significant than those caused by different speech classes and, therefore, recognition degrades dramatically when non-frontal poses are matched against frontal visual models. It is then necessary to develop an effective framework for pose invariant lipreading. In particular, we are interested in pose-invariant methods which can easily be incorporated in the AV-ASR systems developed so far for ideal frontal conditions. In fact, the same problem exists in the face recognition task and it is natural to apply the methods adopted in that field to the lipreading problem. We thus propose to introduce a pose normalization step in a system designed for frontal views, that is, we generate virtual frontal views from the non-frontal images and rely on the existing frontal visual models to recognize speech. The pose normalization block has also an effect on the fusion strategy, where the weight given to the visual stream should reflect its reliability. We can expect that the virtual frontal features generated by the pose normalizer from lateral views will be less reliable than the features extracted directly from frontal images and, therefore, the weight assigned to the pose-normalized visual stream on the audio-visual classifier should account for it.

Previous work on this topic is limited to Lucey et al [6–8], who projected the visual speech features of complete profile images to a frontal viewpoint with a linear transform. We introduce other projection techniques applied in face recognition to the lipreading task and justify their use on the different feature spaces involved in the lipreading system: the images themselves, a smooth and compact representation of the images in the frequency domain or the final features used in the classifier. The effectiveness of the different

methods is supported by lipreading experiments and the effects of the pose normalization in the audio-visual fusion strategy analysed with AV-ASR experiments. The main contributions of this work, which have been partially presented in our conference paper [9], are the adaptation of pose-invariant methods used in face recognition to the lipreading system and the study of its effects on the weight associated to the visual stream in the classifier.

The paper is organized as follows. Section  reviews the structure of an AV-ASR system and explains how the pose-invariance is introduced. In section  we first present the techniques adopted in face recognition to obtain a multi-pose system, adapt some of them to the lipreading problem and study the different feature spaces where the pose-normalization can take place. Finally, experimental results are reported in section  for visual and audio-visual systems and conclusions are drawn in section .

## Audio-Visual Speech Recognition

In terms of the visual modality, AV-ASR systems differ in three major aspects: the visual front-end, the audio-visual integration strategy and the pattern classifier associated to the speech recognition task. In figure **??**, we present a typical AVSR system. First the audio front-end extracts the audio features that will be used in the classifier. This block is identical to that of an audio-only ASR system and the features most commonly used are perceptual linear predictive [10] or Mel frequency cepstral coefficients [11, 12]. In parallel, the face of the speaker has to be localized from the video sequence and the region of the mouth detected and normalized before relevant features can be extracted. Typically, both audio and visual features are extended to include some temporal information of the speech process. Then, the feature streams are feed to statistical classifiers, where usually a hidden Markov models (HMM) [13] is used to estimate the most likely sequence of phonemes or words. The fusion of information between modalities can happen at three stages: merging the extracted features before going through pattern classifiers, on the statistical models used for the likelihood computations or taking the decisions on the pattern classifier. In the following we focus on the visual modality, in particular in the blocks affected by the pose changes on the speaker: the extraction of visual features from images of the mouth and the integration of the visual and audio streams. Finally, we describe the standard AV-ASR system that we adopt and describe how pose normalization can be included in it.

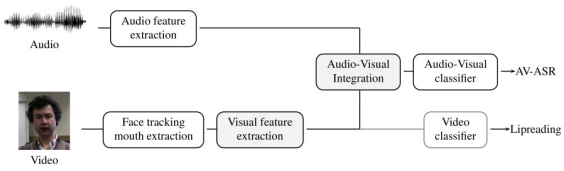



Figure 1: Structure of audio-visual ASR system. Lower part corresponds to the lipreading system. The Audio-Visual system involves also the audio feature extraction and the audio-visual classifier.

### Visual front-end

The first task on the visual front-end is to identify and extract a normalized region of interest (ROI), which is usually a rectangle centred on the mouth of the speaker. The normalization of the region of interest requires a robust face and mouth tracking method as it has to center, align and scale all the images of the mouth in the sequence; making recognition invariant to small movements of the speakers head and to the distance between the speaker and the camera. This step, which is not usually part of the lipreading but the face recognition problem, is critical for the rest of the system and induced the term *front-end effect* to refer to the effects of the ROI extraction in the performance of the speech recognition system.

Two main types of features are used for visual speech recognition [1]: appearance based features extracted directly from the pixels of the ROI and shape based features extracted from the contour of the speaker's lips. Several works [14, 15] report that appearance-based features outperform shape based ones and are, therefore, the features commonly chosen in lipreading and AV-ASR systems. In this approach, the pixels of the ROI themselves are used as features and, consequently, locating the ROI needs to be done with very good precision [16] and the front-end effect carefully considered. The dimensionality of the obtained feature-vector is usually too large to allow an accurate statistical modelling in the classifiers and dimensionality reduction techniques are necessary. The most popular of these techniques are image compressing transforms [17], as principal components analysis or the discrete cosine transform (DCT). They reduce the size of the images by eliminating redundancy, but there is no guarantee that they are appropriate for the classification task. Linear Discriminant Analysis (LDA) [18] is a transform capturing relevant information for classification is and is thus commonly used in AV-ASR. Other supervised transforms based on ideas from information theory have also been proposed for AV-ASR [19–22], but LDA is widely used because it is simple (linear), gives good results and can easily incorporate dynamic information. Dynamic features measure the visual motion during speech and are more robust to skin color or illumination conditions than the original features. This motion can be represented either by delta images or transforms measuring the inter-frame change of the features like, for instance inter-frame LDA.

### Audio-visual integration and classification

Audio-visual integration can be grouped into two categories: feature and decision fusion [1]. In the first case, the audio and visual features are combined projecting them onto an audio-visual feature space, where traditional single-stream classifiers are used [23–26]. Decision fusion, on its turn, processes the streams separately and, at a certain level, combines the outputs of each single-modality classifier. It provides a more flexible set for modality integration, it allows weighting the contribution of each modality in the classification task and is the technique usually adopted. In this case, the audio-visual integration is integrated into the statistical models of the classifier.

In the statistical models used in AV-ASR, the features of the audio and visual streams are assumed class conditionally independent [27, 28], the joint probability distribution is then factorized into single-stream distributions and stream weights $\lambda_A, \lambda_V$ are introduced to control the importance of each modality in the classification task [29, 30]. The resulting joint probability distribution reads

$$p\left(x_A, x_V | q = q_i\right) = p\left(x_A | q = q_i\right)^{\lambda_A} p\left(x_V | q = q_i\right)^{\lambda_V} , \tag{1}$$

where $x_A$, $x_V$ are the audio and visual features and $q$ the class variable. This weighting scheme is naturally introduced in the HMM classifiers by means of multi-stream HMMs [31]. In multi-stream HMMs, independent statistical models like Gaussian mixtures [32] are used to compute the likelihood of each stream independently, which are then combined accordingly to the integration technique. In early integration the streams are assumed to be state synchronous and the likelihoods are combined at state level as indicated by Equation 1. Late integration, in its turn, combines the likelihoods at utterance level, while in intermediate integration the combination takes place at intermediate points of the utterance. The weighting scheme, nonetheless, remains the same and early integration is generally adopted [33]. A common restriction is that the weights $\lambda_A$, $\lambda_V$ sum up to one, which assures that the relation between the emission likelihoods and transition probabilities is kept the same as in single-stream HMMs.

### Our lipreading System

Our speech recognition system is similar to the state-of-the-art presented in [3], which we take as a model system and introduce in it the pose normalization. On the following, we describe our system, giving more details for the blocks which play a role on the pose normalization task.

In order to minimize the front-end effect, we work with sequences where the speaker wears blue lipstick and we can accurately track the mouth by color information. In general, we must rely on a face tracker to find

and track the mouth of the speaker, while in blue-lipstick sequences the position of the lips can be obtained by simple operations in the color space. In our work, we want to study the effects of pose normalization on the visual features and the weighting strategies, not on the face and mouth tracking. Besides, face trackers are usually optimized for frontal poses, are less reliable with non-frontal views and, therefore, the final results of the AV-ASR system can be much affected by the front-end effect. To overcome thi issue, we use sequences where the speaker wears blue lipstick and the mouth ROI position is extracted in the hue domain, detecting for each frame the lips and estimating their size. Afterwards, a sequence of normalized mouths of $64 \times 64$ pixels is extracted and used in the definition of the visual features.

In the second block of our system, state-of-the-art audio and visual features are extracted. In terms of audio features, we adopt Mel Frequency Cepstral Coefficients (MFCC), together with their first and second time derivatives and their means removed by Cepstral Mean Subtraction [34]. For the visual counter-part, we choose appearance-based features and the following sequence of dimensionality reduction transforms. From the original ROI images $x_I$ and $y_I$, we extract a compact low-dimensional representation of the image space retaining only the first 140 DCT coefficients in zig-zag order in $x_F$, $y_F$. To normalize the features for different speakers and sequences, we remove their mean value over the sequence, in an equivalent technique to the Cepstral Mean Subtraction applied to the audio features, and finally, LDA transforms are applied to further reduce the dimensionality of the features and adapt them to the posterior HMM classifier. First, intra-frame LDA reduces to 40 the dimensionality of the features while retaining information about the speech classes of interest, phonemes in our case. Afterwards, inter-frame LDA incorporates dynamic information by concatenating 5 intra-frame LDA vectors over adjacent frames and projecting them via LDA to the final features $x_L$, $y_L$, which have dimension 39 and will be modelled by the HMMs.

The classifiers used are single- and weighted multi-stream HMMs [35]. In the case of AV-ASR, the use of weighted multi-stream HMMs incorporates the audio-visual integration into the classification task, which is done at state level with the weights leading to best performance on an evaluation data.

In our system, see Figure 2, we assume the pose to be known and introduce a pose normalization block to create virtual frontal features from non-frontal ones at different stages of the visual feature extraction. When the transformations are applied directly to the image space, the pose normalization takes place after the mouth extraction, indicated by number 1 in Figure 2. In case of applying the pose-transformation to the selected DCT or LDA features, the transformation block is introduced after the corresponding feature extraction, numbered 2 and 3 in Figure 2.

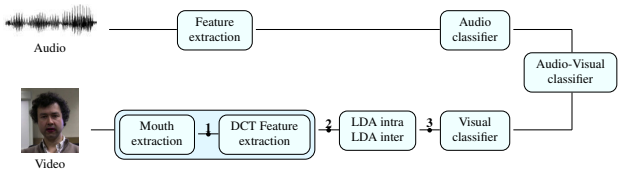



Figure 2: Structure of audio-visual ASR system. Lower part corresponds to the lipreading system. The Audio-Visual system involves also the audio feature extraction and the audio-visual classifier.

## Pose-Invariant lipreading

In this Section we present the techniques adopted in face recognition to obtain a multi-pose system, justify the choice of linear regression (LR) as the technique best suited to our AV-ASR system and study the different feature spaces where the pose-normalization can take place.

### From face recognition to lipreading

The techniques proposed for pose-invariant face recognition can be classified into viewpoint transform and coefficient-based techniques [36]. The coefficient based approach estimates the face under all viewpoints given a single view either by defining pose-invariant features known as "face lightfields" [37] or by estimating some 3-D face model coefficients [38]. In the viewpoint transform approach, on the contrary, the face recognition system is designed and optimized for the dominant view (frontal in our case) and a preprocessing step transforms the input images corresponding to undesired (non-frontal) poses to the desired view [36]. The same two strategies can be applied to the lipreading task. We adopt the viewpoint-transform approach because lipreading predominantly takes place with frontal views and coefficient-based techniques would suffer from over-generalization, i.e., only a small fraction of the time would system benefit from the definition of pose-invariant features, being outperformed most of the time by a system optimized for frontal views.

In the viewpoint transform approach there are two strategies to generate virtual frontal views from non-frontal poses: 3-D models [39, 40] and learning-based methods [41, 42]. In the first case, a 3-D morphable model of the face must be built from 2-D images before virtual views from any viewpoint can be generated with graphic rendering techniques. It is computationally expensive and time consuming to match the input 2-D image with the 3-D model and, therefore, that technique is not aimed to the real-world applications of AV-ASR. To overcome that issue, learning-based approaches learn how to estimate virtual views directly in the 2-D domain, either via a 2-D face model or from the images themselves. Working directly with the face images a simple and yet effective way to project the images from lateral to frontal views is based on linear regression [36, 43] (LR). We choose to rely on the images themselves instead of introducing a mouth model

to estimate the virtual views for several reasons. First, most lipreading systems use directly images of the mouth as visual features and do not require mouth or lip models, which we do not want to introduce in the lipreading system only for the sake of pose normalization [44]. Secondly, the visual features extracted from the images themselves are more informative than features based on lip-modelling, as they include additional information about other speech articulators such as teeth, tongue and jaws also useful in human speech perception [45]. These pose normalization techniques involve transforms that can be quickly computed and allow real-time implementations required in most AV-ASR applications. Finally, appearance based features directly obtained from the image pixels are generic and can be applied to mouths of any viewpoint compared to lip models which have to be developed for any possible view.

**Linear Regression in multi-pose face recognition**

Given a set of $M$ training examples of the undesired viewpoint $Y = \left[y^1 \dots y^M\right]$ and their synchronous examples on the target viewpoint $X = \left[x^1 \dots x^M\right]$, a matrix $W$ performing LR is determined minimizing the cost function $Q$

$$Q(W) = \sum_{i=1}^{M} ||x^i - Wy^i||^2 + \beta||W||^2, \tag{2}$$

which measures the mean square error on the training dataset and might include a Tykhonov regularization term (weighted by parameter $\beta$) introducing additional smoothness properties and leading to a Ridge Regression [46]. The well-known solution to the LR is given by $W = XY^T\left(YY^T + \beta I\right)^{-1}$, with $I$ the identity matrix.

Linear Regression is theoretically justified when images of the same object but from different poses are subject to the same illumination. In the case of face recognition, in [43] the authors show that if the face images are well aligned, there exists an approximate linear mapping $x_I = W^I y_I$ between images of one person captured under variable poses $x_I$ and $y_I$, which is consistent through different people. Unfortunately, in real-world systems face images are only coarsely aligned, occlusions derived from the 3-D nature of faces affect the different views and the linear mapping assumption no longer holds. To this end, the authors propose the use of a piecewise linear function to approximate the non-linear mapping existing between images from different poses. The main idea of the proposed method lies in the intuitive observation that, by partitioning the whole face into multiple patches, linearity of the mapping for each patch holds since the face misalignment and variability between different persons is reduced. They call that technique local LR (LLR) in opposition to the previous implementation of LR, which considered the images as a whole and is therefore designated as global LR (GLR) and can be considered a particular case of LLR with only one patch.

(a) Frontal patch (b) Lateral patch

Figure 3: Frontal and corresponding lateral patch definition for the LLR computation

Intuitively, LLR partitions the whole non-frontal image into multiple patches and applies linear regression to each patch. Given the training set $\{X, Y\}$, each face image is divided into blocks of rectangular patches $\{X_i, Y_i\}_{i=1...N}$. Then, for each pair of frontal and lateral patches the linear regression matrix $W_i$ is computed as in the GLR case. In the testing stage, given an input image with known pose, it is partitioned it into patches and used to predict each frontal patch with the corresponding matrix $W_i$. Afterwards, all the virtual frontal patches are combined into a whole vector to construct a virtual frontal image. For the frontal views a uniform partition of the images is adopted, while for non-frontal images each patch contains surface points of the same semantics as those in the corresponding frontal patch. In the case of a completely profile image, for instance, we associate two frontal patches to each profile one imposing symmetry to the frontal view. See Figure 3 for a pair-example of patches defined across different views. The patches can be adjacent or overlap, alleviating in that case the block effect but increasing the cost of reconstruction as the value associated to a pixel sampled by several patches is then computed as the mean of the specific pixels in the overlapping patches. Consequently, the patch size and overlapping are parameters to choose for the LLR method to succeed. While a too large patch size suffers from the linear assumption and can lead to blurring of the images, a patch too small is more sensible to misalignments and produces artefacts on the reconstructed image. The overlapping criteria, on its turn, is a trade-off between over-smoothing (high overlapping of patches) and introducing block effects on the reconstructed images (adjacent patches).

#### Linear Regression and lipreading

In our work, the LR techniques are applied considering $X$ and $Y$ to be either directly the images from frontal and lateral views $X_I$,$Y_I$ or the visual features extracted from them at different stages of the feature extraction process. A first set of features $X_F$,$Y_F$ are designed to smooth the images and obtain a more compact and low-dimensional representation in the frequency domain. Afterwards, those features are transformed and their dimensionality anew reduced in order to contain only information relevant for speech classification,

leading to the vectors $X_L$,$Y_L$ used in the posterior speech classifier.

The visual features $X_F$,$Y_F$ are the first coefficients of the two-dimensional DCT of the image following the zigzag order, which provide a smooth, compact and low dimensional representation of the mouth. Note that the selected DCT can be obtained as a linear transform, $X_F = SDX_I$, with $D$ the matrix of two dimensional DCT basis transform and $S$ a matrix selecting the DCT coefficients of interest. Therefore there is also an approximate linear mapping $DW^I D^T$ between the DCT coefficients of the frontal and lateral images $x_D = Dx_I$, $y_D = Dy_I$. Indeed, as the DCT forms an orthonormal base in the image space, we can write the original linear mapping between the images as

$$x_D = Dx_I = DW^I y_I = DW^I \left(D^T D\right) x_I = DW^I D^T y_D. \tag{3}$$

Consequently, if all DCT coefficients are selected and $S = I$, the DCT coefficients obtained from $W^I$ by projecting images $y_I$ and the $W^F$ projected coefficients from $y_F$ coincide. The linear relationship, however, no longer holds when we consider only a reduced set of DCT coefficients $x_F = SDx_I$ and the transform $W^F$ found with the LR method should be considered an approximation of the non-linear mapping existing between any pair of reduced DCT coefficients $x_F$ and $y_F$. In that case, selecting the DCT features corresponding to lower frequencies to compute the transform $W^F$ corresponds to smoothing the images previous to the projection and estimating a linear transform forcing the projected virtual image to be smooth by having only low-frequency components. Moreover, the lower-dimensionality of $X_F$, $Y_F$ compared to $X_I$, $Y_I$ improves accuracy of the LR matrix estimation due to the *Curse of Dimensionality* [47], which states that the number of samples necessary to estimate a vectorial parameter grows exponentially with its dimensionality. In that sense, the effect of the regularization parameter $\beta$ is more important in the estimation of $W^I$ than $W^F$, as imposing smoothness reduces the number of required samples.

It is important to note that the proposed LLR technique on the DCT features provides a different meaning to the patches. If we choose the patches to be adjacent blocks of the DCT coefficients, we are considering different transforms for different frequency components of the image. Consequently, we use an equal partition of the selected DCT coefficients to define the frontal and associated lateral patches in the LLR transform. In that case, LLR approximates the existing non-linear mapping between frequency features $X_F$ and $Y_F$ by distinct linear functions between the different frequency bands of the images.

Another option to apply pose normalization, is to project the final features $X_L$, $Y_L$ used in the pattern classifier. Those features are obtained from linear dimensionality reduction transforms aimed at speech classification [44]. The transforms are usually based on LDA, which is a supervised transform projecting the

DCT features $x_D$, $y_D$ to the linear subspace maximizing the separability of the $C$ speech classes. Specifically, LDA finds the $K$-dimensional linear subspace maximizing the projected ratio $R = S_w^{-1} S_b$ between the between-class scatter matrix $S_b$ and within-class scatter matrix $S_w$, defined as

$$S_w = \sum_{i=1}^{C} p_i \Sigma_i \qquad S_b = \sum_{i=1}^{C} p_i \left(\mu - \mu_i\right) \left(\mu - \mu_i\right)^T, \tag{4}$$

where $p_i$ is the percentage of samples on the training set belonging to the class $i$, $\mu_i$ and $\Sigma_i$ are the mean and covariance matrix for those samples and $\mu$ is the mean value of all the training samples in the dataset. The LDA projection matrix is then defined by the eigenvectors of $R$ with $K$ largest associated eigenvalues. If there is a linear mapping between the original features $x = Wy$, we can also relate the corresponding LDA projections with a linear mapping. Observing that

$$S_b^x = W S_b^y W^T \qquad S_w^x = W S_w^y W^T \tag{5}$$

it is easy to prove that if $v$ is an eigenvector of $R^y$ with eigenvalue $\lambda_v$, then $W^{-1}v$ is an eigenvector of $R^x$ with the same eigenvalue and, consequently, there is also a linear mapping between the LDA projections associated to the frontal and lateral views. Two extra considerations have to be taken into account for the projection of the $X_L$ and $Y_L$ features. First, $X_L$ and $Y_L$ are obtained by applying LDA into the reduced DCT features $X_F$ and $Y_F$, which means that the projection by $W^L$ is only a linear approximation of the real mapping between the LDA features in the same way $W^F$ linearly approximates the relation between $X_F$ and $Y_F$. Second, two stages of LDA are needed to obtain $X_L$ and $Y_L$ from $X_F$ and $Y_F$, a first intra-frame LDA and then an inter-frame LDA on concatenated adjacent vectors extracted from the intra-frame LDA. In the intra-frame LDA, $x = x_F$, $y = y_F$ and $W = W^F$ in Equation 5, from which we obtain LDA projected vectors $x_l$ and $y_l$, related with an approximated linear mapping $W^l$. In the inter-frame LDA, each $x$ and $y$ corresponds to the concatenation of 5 time-adjacent vectors $x_l$ and $y_l$, and thus the approximated linear mapping $W$ is given by a block matrix whose diagonal entries correspond to 5 block matrices $W^l$. As a consequence, if the linear approximation of $X_F = W^F Y_F$ holds, then it is also a valid assumption for the projection of the speech features by $X_L = W^L Y_L$.

The performance of the linear regression applied to the images or the extracted features can be analysed by the cost function $Q$ normalized to the size of the vectors $X$ and $Y$. The mean value taken by the cost function in our training dataset is presented in Figure 4, where we observe the effect of the curse of dimensionality as the error associated to the estimation of $W^I$ are considerably higher than for $W^F$ or $W^L$. As experiments will show, the smaller dimensionality of the DCT and LDA features allow us to learn more

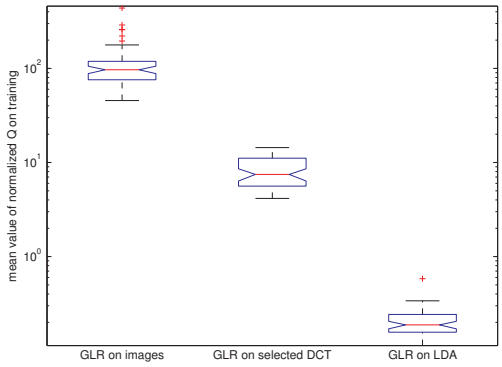


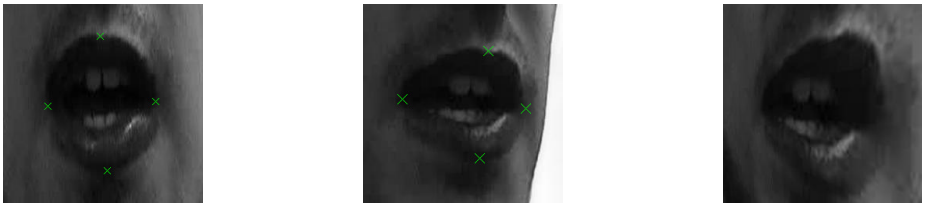

Figure 4: Value of cost function for the LR training sequences applied to the images $X_I$, the selected DCT coefficients $X_F$ and LDA speech features $X_L$

accurately the GLR transform matrices, leading also to better speech recognition performance of the GLR technique in those feature spaces.

Consideration should be given to the fact that applying the pose normalization on the original images, or even to the low-frequency DCT coefficients, is independent of the features we posteriorly use for speech recognition and could be adopted with other appearance or contour-based visual speech features. The use of the LDA features, however, is specific to the speech recognition system and involves an additional training of LDA projections for the different poses. In that sense, applying the LR techniques to the original images provides a more general strategy for the multipose problem, while the LDA features might be able to exploit their specificity for the speech recognition task.

**Projective transforms on the images**

A simple option when working with the images themselves is to estimate a projective transform from the lateral to the frontal views by as a change of the coordinate systems between the images. In fact, as the difference in poses involves an extra dimension not taken into account in the projective model (3-D nature of the head rotation), that approach can only be justified for small pose changes, being impossible to implement, for instance, for 90° of head rotation. Nevertheless, we include it in our experiments for comparison purposes. In that case, we learn a $3 \times 3$ projective transform $T$ between the image coordinates in a semi-manual and automatic procedure. The coordinate points $P$ used for that purpose are the corners of the lips, the center of the cupid's bow and the center of the lower lip contour for the different poses. In the manual procedure, we selected several frames of each sequence, manually marked the position of those 4 points for the frontal and lateral views and estimated the transform $T$ minimizing the error of $P_{frontal} = TP_{lateral}$ over the selected



(a) Frontal image (b) Lateral image (c) Projected image

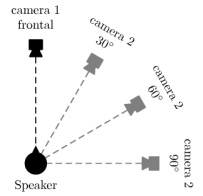
Figure 5: Original frontal (a) and lateral (b) view images with manually annotated points $P_{frontal}$ and $P_{lateral}$. (c) Virtual frontal view image obtained by the estimated projective transform $T$.

frames of the sequence. For the automatic method, we used the color information (the speakers wear blue lipstick) to detect the lip's contour and the position of the points $P$ used to estimate the projective transform. Examples of the images obtained with that method are shown in Figure 5, where the deformations caused by neglecting the 3-D nature of head rotation are obvious. That effect is not encountered with the LR technique applied to the images, as the training stage on the process is responsible of learning how the mouth views change with the poses. In that sense, the projective techniques can be used with any kind of images and do no exploit the fact that all the images correspond to mouths.

## Experimental Results

We present two sets of experiments: one on lipreading studying the adaptation of the visual stream to multi-pose conditions and another on AV-ASR analysing the effects of the pose normalization on the audio-visual integration strategy. In lipreading experiments we first quantify the loss of performance associated to non-frontal poses in ASR, we then justify quantitatively the necessity of a pose normalization step and then analyze the performance of the pose normalization strategies proposed in Section . In audio-visual experiments, the first question to answer is whether the visual stream can still be exploited for speech recognition after the pose normalization has taken place, something that previous works [4, 5] studyig AV-ASR in realistic working conditions failed to do. In AV-ASR we are also interested in the influence of the pose-normalization in the final performance and, specially, on the optimal value of the weight associated to the visual stream.

The technical details of the experimental set-up are the following. The task considered is connected speech recognition under different speaker poses relative to the camera. Training and testing has been done with the multi-speaker paradigm (all speakers are on train and test set but with different sequences) with three fold cross-validation and the results are given in terms of word accuracy. The same multi-speaker




(a) Camera positions

(b) Speaker frontal pose

(c) Speaker pose 60°

Figure 6: Schema of simultaneous recordings with different poses and sample images from the database

cross-validation is used to estimate the LR transforms for the different poses and features. The parameters of the audio and visual feature extraction blocks and classifiers are chosen based on experiments with an evaluation dataset to optimize speech recognition. To fairly analyze the performance associated to frontal and lateral views, the same kind of classifiers are trained for each possible pose: frontal (abbreviated as F-classifier) and lateral at 30°, 60° and 90° of head rotation (L-classifiers). The HTK tool-kit [48] is used to implement three-state phoneme HMMs with a mixture of three Gaussians per state. For the multi-stream HMMs, the same number of states and Gaussians than in single-stream case is used. The parameters of the model are initialized with the values estimated for independent audio and visual HMMs and posteriorly re-estimated jointly. The audio and visual weights are considered fixed parameters of the system, restricted to sum up to one and chosen on an evaluation dataset to optimize speech recognition.

### Database

For our experiments we required speech recordings with constrained non-ideal visual conditions, namely, fixed known poses and natural lighting. To that purpose we recorded our own database, which is publicly available at our webpage. It consists of recordings of 20 native french speakers with simultaneous different views, one always frontal to the speaker and the other with different lateral poses.

The recordings involve one frontal camera plus one camera rotated 30°, 60° and 90° relative to the speaker in order to obtain two simultaneous views of each sequence, see figure 6. The first camera was fixed with a frontal view, while the second camera provided different lateral views. For each possible position of the second camera, the speaker repeated three times the digits, giving a total of $3 \times 3$ couples of repetitions of each digit: 9 for frontal views and 3 laterals at 30°, 60° and 90° of head rotation. To comply with the natural conditions, the corpus was recorded without paying much attention to the lighting conditions, which resulted in shadows on some images under the nose and mouth of the subjects. The videos were recorded with two high-definition cameras CANON VIXIA HG20, providing $1920 \times 1080$ pixels resolution at 25 frames

per second, and included the head and shoulders of the speaker.

In terms of audio set-up, two different micros were used for the recordings, an external micro close to the speaker's mouth, without occluding its view, and the built-in micro of the second camera. That set-up provided two conditions for the audio signal, a clean audio signal obtained with the external microphone (directional dynamic micro tailored for human voice, Sony F-V120) and a naturally noisy one due to the use of a more generic microphone and its distance to the speaker. Audio was recorded with a sample rate of 48000 Hz and 256 kbps for both micros, with the external micro connected to the static camera and the built-in micro of the second camera used both to obtain natural noisy audio data and to synchronize the videos. Indeed, both videos were synchronized based on the audio signal because it offered better time resolution than a pairing of the video frames (video frame resolution equalling 40 milliseconds). Thus, for the two audio signals we computed the correlation of their normalized MFCC features within each manually segmented word, obtained an estimate of the a delay for each word and averaged over the whole sequence. The same delay was considered for the video signals, after correcting for the difference in distance between the two micros and the speaker.

The word labelling of the sequences was done manually at the millisecond and phone labels were posteriorly obtained by force alignment of the clean audio signals with the known transcriptions.

#### Visual Speech Recognition

In a first set on experiments we quantify the loss of performance of a lateral system compared to a fully frontal one. To that purpose, we paired the frontal and lateral sequences and test each sequence with the corresponding classifier, i.e F-sequences with F-classifier and, for each possible head rotation, the L-sequences with their L-classifier. That gives us a measure of how visual speech degrades with the different poses, presented in Figure 7. As happens with human lipreading [49], speech recognition deteriorates with non-frontal speaker poses, which of course is more acute for 90° (9% of loss of performance with respect to the frontal system) than for 60° (5% of loss of performance with respect to the frontal system). We also present in Figure 7 the performance of the F-classifier tested with the L-sequences when no pose normalization is applied, i.e., there is a mismatch on the train/test conditions in terms of pose and so the system performs poorly, with mean word accuracy dropping from 71% to 22%. This justifies the necessity of pose normalization.

As we have already explained, however, our objective is to adapt the extracted visual features and use a classifier optimized for frontal poses, so that on the following we will test the different pose normalization

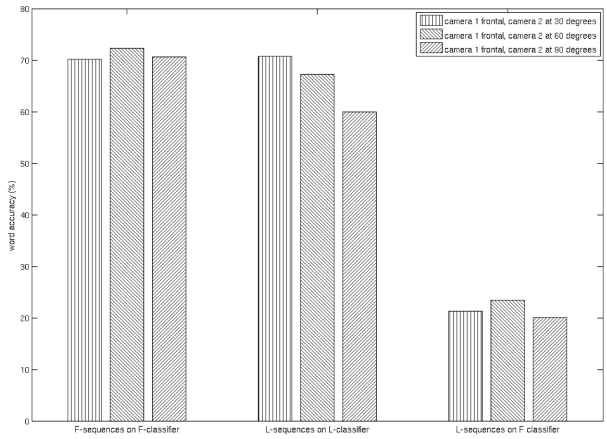



Figure 7: Loss of performance due to non-frontal poses and train-test pose mismatch

techniques with the L-sequences on the classifier trained and optimized for frontal sequences. In that sense, we should not only compare the results of the pose-normalized L-sequences to the corresponding F-sequences with the F-classifier, but also to the performance of the lateral views when tested on their L-classifier. The results of F-sequences on F-classifier represent the best we can do in terms of original pose and trained system, while the results of L-sequences on L-classifier represent the best we can do when the original images present a non-frontal pose with a lipreading system adapted to it.

The performance results are shown in Figure 8. For each possible feature space, we choose the best-performing LR technique: LLR on the images (split in 32x32 pixel patches with 75% overlapping, $\beta = 15$) and GLR on the selected DCT ($\beta = 5$) and LDA ($\beta = 0$). As expected, the features obtained after the pose normalization can neither beat the schema F-sequences on F-classifier, because there is a loss of valuable information in the non-frontal images, nor obtain the performance of L-sequences on L-classifier, due to the limitations of the pose normalization techniques. For the different poses, the projected LDA features clearly outperform the other techniques (between 4% to 12% of loss of accuracy for the different poses compared to F sequences), making use of the specificity of the features for speech recognition compared to the more general image or DCT feature spaces (accuracy loss 25% to 34% compared to the frontal views). The fact that the original images and the selected DCT coefficients present similar performance with different LR techniques and regularization parameter $\beta$ is justified by the LR training stage and the effects of misalignment on the images. The curse of dimensionality states that, with a limited amount of training data, we are only able to accurately estimate the values of the $LR$ transform up to a certain dimensionality. Consequently, the LLR technique applied to the images outperforms the GLR not only because it reduces the effects of misalignment

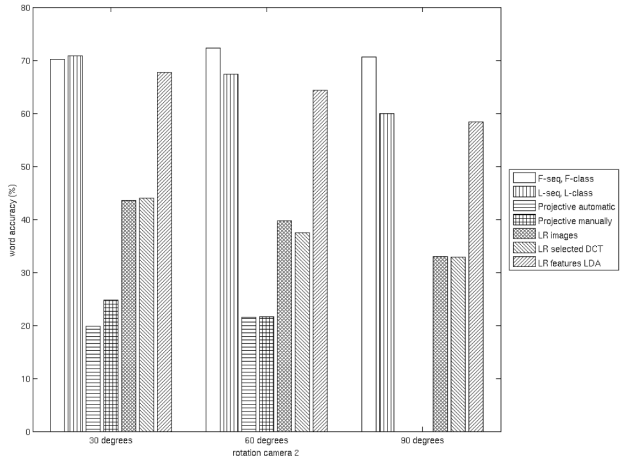



Figure 8: Performance of the C-classifier on projected L-sequences

on the images, but also because it can more accurately estimate the values of the linear transforms in a feature space of the size of the patches instead of the image. Note too that, as expected, the projective transforms obtain poor recognition results (50% of accuracy loss compared to the frontal views) as they neglect the effects of 3-D pose changes on the mouth views.

Comparing the different LR techniques applied to the original images, see Figure 9, we see that LLR performs better than GLR. Splitting the images in 4 patches (half the height and half the width of the original image, which is designated as $N = 4$ in the figure) and allowing an overlapping of 75% of the patches ($o = 3/4$) lead to the best results, showing the expected trade-off between the size-of the patch and the overlapping. A patch size too large (GLR case) suffers from the linear assumption and leads to blurring of the images, while a patch too small ($N = 16$, for instance) is more sensible to misalignments. Similarly, for each patch size, a high overlapping of patches ($o = 7/8$ for $N = 4$ or $o = 3/4$ for $N = 16$) results in over-smoothing, while low values cause block effects on the reconstructed images ($o = 0$, no overlapping). At the same time, the value of the regularization parameter $\beta$ leading to best results increased with the size of the patches.

For the selected DCT coefficients, however, the general mapping defined by GLR obtains better results, while for the LDA case both techniques perform similarly, see Figure 10 and Figure 11. The worst performance of the LLR with the selected DCT features can be explained by the observation that the patches defined in the DCT space correspond to high and low-frequency components of the images. It seems likely, therefore, that a linear transform between the low-frequency components of the images exist, but that assumption does not hold for the high-frequency components associated to image details. In the case of LDA

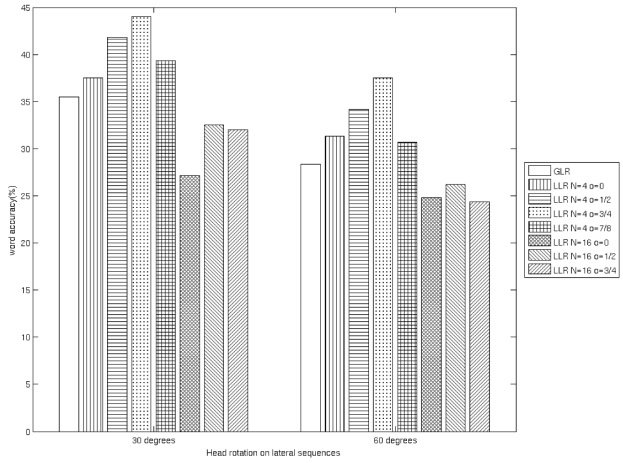



Figure 9: Performance of the C-classifier on projected L-sequences in image space

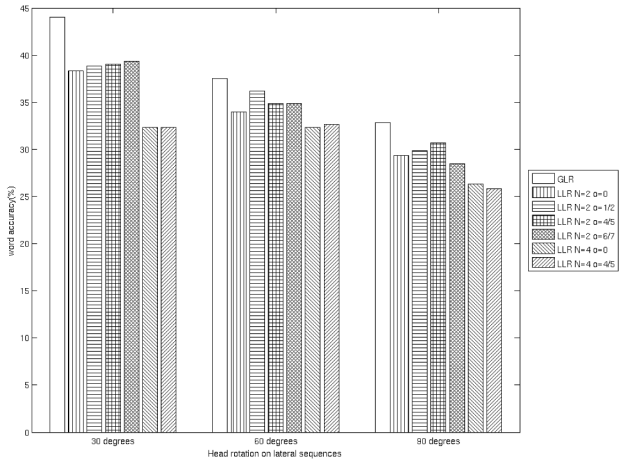



Figure 10: Performance of the C-classifier on projected L-sequences in selected DCT space

features there is no interpretation of the patches defined on the LLR technique[1] and we observe that the mapping between the frontal and lateral features can be similarly approximated wit the GLR and LLR techniques. In fact, there is no statistical difference between the performance of the GLR and LLR techniques in that feature space.

We can relate the performance of the LR techniques to the values taken by the normalized cost function $Q$ on the testing data. In Figure 12 we show the values taken by the $Q$ function of the GLR technique for 20 testing sequences against the speech recognition performance those same sequences obtained. We observe

[1]For simple LDA we can interpret the patches as directions on the original space maximizing the projected ratio $R$, so that if we sort the eigenvectors on the LDA projection according to their eigenvalue, we could interpret the patches as linear subspaces decreasingly maximizing the projected ratio. However, as we include intra and inter-frame LDA in the $W^L$ transform, no interpretation is possible for the patch definition on the $x_L$, $y_L$ space.

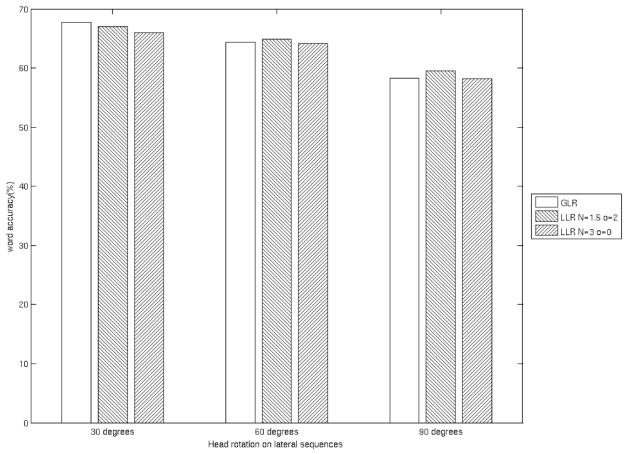



Figure 11: Performance of the C-classifier on projected L-sequences LDA feature space

three separated groups on the $x$ and $y$-axis associated to the different feature spaces. The LDA features obtained better performance and presented lower values of the cost error function $Q$ associated to the GLR the projection, while the original images and the selected DCT coefficients show higher values of $Q$ and lower associated performances. As we have already observed, the results with the projected LDA features are more conclusive than with the projected DCT or images, whose difference on the values taken by $Q$ are not reflected on the speech recognition accuracy they obtain. Indeed, $Q$ values are considerably higher for the images, but their performance is only slightly worse than the selected DCT coefficients. It is not worth, thus, to work on the high-dimensional image with the local version of LR instead of using GLR on the reduced DCT space. In that sense, any improvement on the virtual views obtained in the LLR projection of images is lost on their posterior projection to the reduced DCT space. Looking at the dispersion of the $Q$ values within a feature space, we do not observe a correlation with their speech recognition performance. The value taken by the cost function $Q$ can thus be used to select the feature space for the LR projection, but does not allow to estimate the performance of different projected sequences within a feature space.

#### Audio-Visual Speech Recognition

We study how the pose variations influence audio-visual ASR systems. Since the visual stream is most useful when the audio signal is corrupted, we report audio-visual experiments with a noisy audio signal and compare it to an audio-only ASR system. To that purpose we artificially added babble noise extracted from the NOISEX [50] database to the clean audio signal with 7 dB and 0 dB of Signal-to-Noise Ratio (SNR). The audio HMM parameters were trained on clean audio data, but the corrupted signals were used for testing.

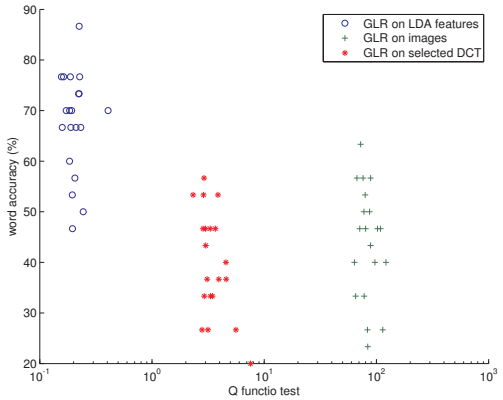
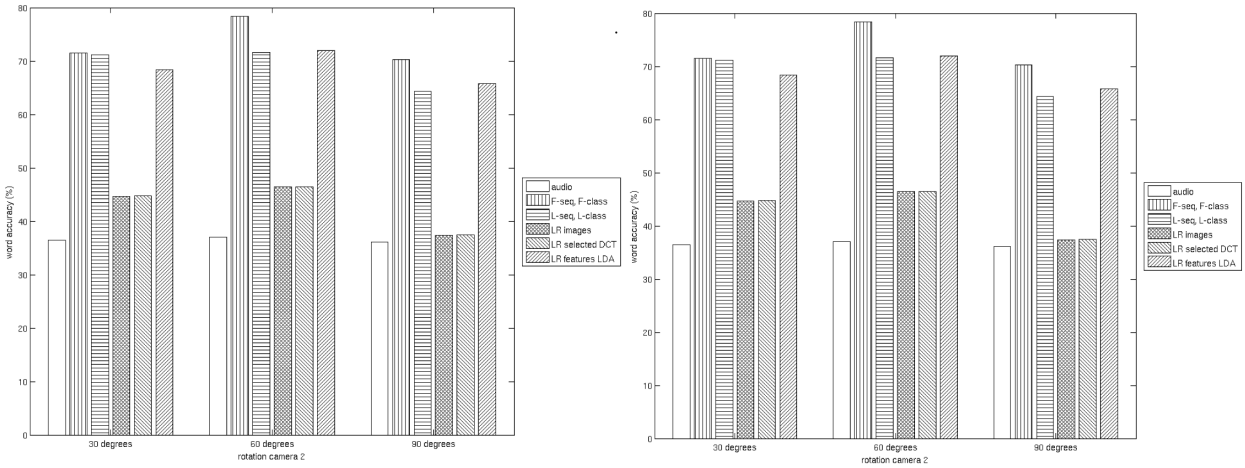



Figure 12: Scatter between the values of the mean square error on GLR projection in the feature space and the speech recognition performance obtained with those features

Figure 13 show the performance for the audio-visual system for frontal and lateral poses. The lipreading block of the audio-visual system correspond to the same sequences and classifiers used in lipreading experiments. The performance of the different streams is coherent with the visual-only experiments, with frontal views outperforming lateral ones and GLR on the LDA space clearly improving upon the other pose normalization methods. Note that the absolute difference in performance between the different visual streams is now reduced. In an audio-visual system, the weight assigned to the visual stream controls to which extend the classifier's decision is based on the visual features and, therefore, differences between visual streams are more evident when the weight assigned to the video is high. The extreme cases correspond to a completely corrupted visual stream, where $\lambda_A = 1$, $\lambda_V = 0$ and the different pose normalization techniques obtain the same performance, and to a corrupted audio signal leading to optimal weights $\lambda_A = 0$, $\lambda_V = 1$ and the performance we already observed in lipreading experiments. Consequently, the differences in performance of the pose normalization methods are more acute with 0 dB than 7 dB of audio SNR. Observe, for instance, how at 7 dB the LR technique applied to the LDA features gives the same performance than the original L-sequences with a L-classifier, but for different values of the video weight. Notice also that the LR projection techniques applied to the original images or the selected DCT coefficients are only able to improve audio recognition when the audio signal is highly corrupted (0 dB), while the projection on the LDA space always ameliorates the recognition of the audio system. The LR results for the images and DCT coefficients at 7 dB point out the fact those techniques are not useful for speech recognition and only increase the confusion of the audio classifier (an audio-visual system outperforms an audio one only when the errors incurred in the audio domain are uncorrelated with the errors in the visual domain, which is not the case here).




(a) 7 dB audio SNR

(b) 0 dB audio SNR

Figure 13: Word accuracy for audio and audio-visual systems with different visual streams and classifiers

We also analyse the value of the video weight $\lambda_V$ assigned to the different sequences and pose normalization techniques and relate it to their performance in lipreading experiments. The weights assigned to the visual stream are presented in Figure 14, where we observe that, as expected, the weight given to the visual modality decreases with the quality associated to the visual stream. For the frontal view sequences $\lambda_V$ takes higher values than for the lateral ones. Similarly, the projected L-sequences with the L-classifier have higher weights than the pose-normalized L-sequences when tested on a frontal classifier and the values for 90° of head rotation are lower than for 30°. We have also found a nice correlation between the values of the optimal visual weight and the stream's performance in lipreading experiments, as presented in Figure 15. We can thus relate the audio-visual and lipreading systems in terms of their performance and visual weights also for non-frontal views and pose normalization purposes.

### Statistical significance of the results

In our experiments we compare different views from the speaker and pose normalization strategies learned and tested on the same data and the results, therefore, reflect differences between the views and pose normalization strategies rather than differences in the test datasets. In this case, the statistical significance of the results cannot be evaluated by means of confidence intervals associated to the performance of each method independently, but requires the comparison of the different methods in a one-to-one basis for the same sentences, speakers and train/test datasets.

In speech recognition a small modification to a system will alter the recognition results in a few sentences

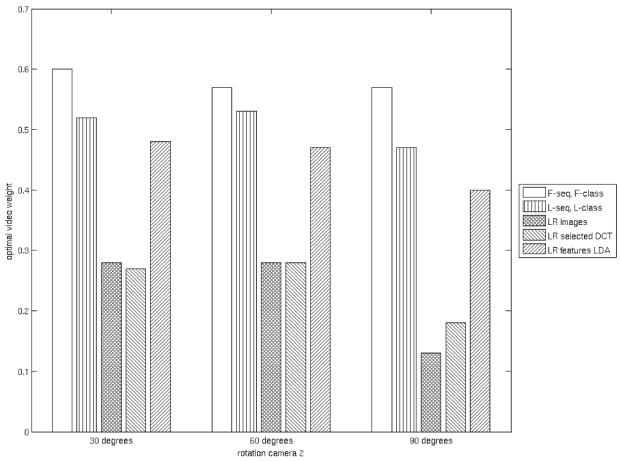



Figure 14: Optimal video weight in the multi-stream system for a corrupted audio with SNR 7 dB

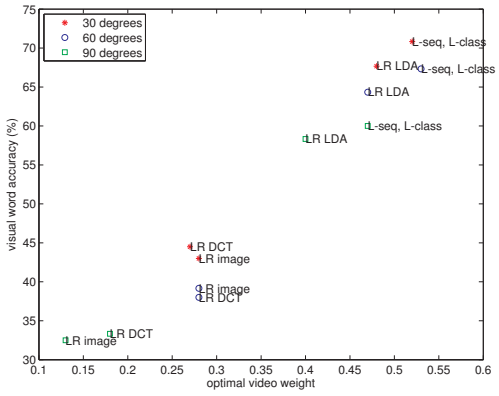



Figure 15: Scatter between optimal video weight and visual-only speech recognition performance for a corrupted audio with SNR 7 dB

or speakers only. Intuitively we would acknowledge a 10% probability of reducing the errors if the number of errors drops on 10% of the sentences while the others remain unchanged. On the other hand, an overall improvement of the word error rate should be considered random if 50% of the sentences improved while 50% degraded. In this work, we use the "probability of error reduction" $p_e$ presented in [51] to assess the differences in performance of the proposed weighting schemes. We refer the reader to the original paper [51] for a detailed description of $p_e$ and present here only the main ideas. Intuitively, we measure the probability of error reduction $p_e$ between two systems A and B counting the number of independent testing samples (sentences in our experiments) that favour system A over B while leaving the rest of the samples unchanged. Formally, however, the computation of $p_e$ requires the estimation of the probability distribution associated to the paired comparison of the systems. To that purpose, we bootstrap the WER obtained by the different weighting methods for independent samples, count the number of samples favouring each system and perform a paired hypothesis test to obtain $p_e$. Bootstraping allows us to estimate the unknown distributions associated to the WER and the paired comparison of the systems and obtain an estimate of $p_e$ which does not depend on the number of sentences used in each comparison and is defined in the same terms used to evaluate speech recognition systems. On the following, only the values of $p_e$ relevant to assess if one method significantly outperforms another are given.

For instance, to compare the loss of performance associated to non-frontal poses, we compute the probability of error reduction obtained when using a frontal system (F-sequences on F-classifier) against a lateral one (corresponding L-sequences on L-classifier). For the sequences where camera 2 presents a view of 30° of head rotation $p_e = 0.52$, that is, lipreading can be performed at 30° of head rotation as the performance degrades on 52% of the cases and improves on 48%. On the other hand, for the 60° and 90° sequences, $p_e$ takes values 0.85 and 0.98 and speech recognition degrades through the majority of training/testing sets.

For the different pose normalization techniques, we compute $p_e$ with respect to a lateral system, that is, the original L-sequences tested on the L-classifier. For the image and DCT feature spaces, performance degrades in every single test case for all the possible lateral views ($p_e = 1$). In the case of LDA feature space (with the GLR technique), performance degrades in 70% of the cases for 30° of head rotation and in 80% for the rest of the lateral views. This analysis shows that, even though the final accuracy of the L-sequences with the L-classifier might be close to the projected LDA features with the F-classifier, there is a significant loss of performance due to the pose normalization.

For the audio-visual experiments, we compare each of the systems to an audio-only recognizer. As pointed out by Figure 13, only the pose normalization applied in the LDA space is able to exploit the visual stream in

the AV-ASR system with 7 dB of SNR, with performance improving in 98%, 95% and 89% of the sequences at 30, 60 and 90° of head rotation in comparison to an audio-only system. This percentage is inferior to 16% and 13% for the DCT or image space, pointing out the fact that the introduction of pose normalization in these feature spaces fails to exploit the visual modality in an AV-ASR system. In a more noisy environment with 0 dB of SNR, the projection on the LDA space is always beneficial, while the DCT and image spaces only do better than an audio-only system in 80% of the cases.

## Conclusions

In this paper we presented a lipreading system able to recognize speech from different views of the speaker. Inspired by pose-invariant face recognition studies, we introduce a pose normalization block in a standard system and generate virtual frontal views from non-frontal images. In particular, we use linear regression to project the features associated to different poses at different stages of the lipreading system: the images themselves, a low-dimensional and compact representation of the images in the frequency domain or the final LDA features used for classification. Our experiments show that the pose normalization is more successful when applied directly to the LDA features used in the classifier, while the projection of more general features like the images or their low-frequency representation fails because of misalignments on the training data and errors on the estimation of the transforms.

In terms of AV-ASR, we study the effects of pose normalization in the fusion strategy of the audio and visual modalities. We evaluate the effects of pose normalization on the weigh associated to the visual stream and analyse for which one of the proposed techniques the audio-visual system is able to exploit its visual modality. We show that only the projection of the LDA features used in the classifier is really able to normalize the visual stream to a virtual frontal pose and enhance the performance of the audio system in noisy environments. As expected, there is a direct relation between the optimal weight associated to the pose normalized visual stream and its performance in lipreading experiments. Consequently, we can simply study the effects of pose normalization in the visual domain and transfer the improvements into the audio-visual task by adapting the weight associated to the visual stream.

## Acknowledgements

This work is supported by the Swiss SNF grant number 200021-130152. ...

## References


1. Potamianos G, Neti C, Luettin J, Matthews I: **Audio-visual automatic speech recognition: an overview**. In *Issues in audio-visual speech processing*. Edited by Bailly G, Vatikiotis-Bateson E, Perrier P, MIT Press 2004.
2. Dupont S, Luettin J: **Audio-visual speech modeling for continuous speech recognition**. In *IEEE Transactions on Multimedia*, *Volume 2* 2000:141–151.
3. Potamianos G, Neti C, Gravier G, Garg A, Senior A: **Recent advances in the automatic recognition of audiovisual speech**. *Proceedings of the IEEE* 2003, **91**(9):1306 – 1326.
4. Potamianos G, Neti C: **Audio-visual speech recognition in challenging environments**. In *Eighth European Conference on Speech Communication and Technology* 2003.
5. Livescu K, Cetin O, Hasegawa-Johnson M, King S, Bartels C, Borges N, Kantor A, Lal P, Yung L, Bezman A, et al.: **Articulatory feature-based methods for acoustic and audio-visual speech recognition**. In *Final Workshop Report, Center for Language and Speech Processing, John Hopkins University*, *Volume 4* 2006.
6. Lucey P, Sridharan S, Dean D: **Continuous Pose-Invariant Lipreading**. In *Proceedings of Interspeech* 2008.
7. Lucey P, Potamianos G, Sridharan S: **A unified approach to multi-pose audio-visual ASR**. In *Interspeech Proceedings* 2007:650–653.
8. Lucey P, Potamianos G, Sridharan S: **An Extended Pose-Invariant Lipreading System**. In *International Workshop on Auditory-Visual Speech Processing* 2007.
9. Estellers V, Thiran JP: **Multipose Audio-Visual Speech Recognition**. In *EUSIPCO Proceedings* 2011.
10. Hermansky H: **Perceptual Linear Predictive (PLP) Analysis of Speech**. *Journal of the Acoustical Society of America* 1990, **87**(4):1738–1752.
11. Mermelstein P: **Distance measures for speech recognition, psychological and instrumental**. *Pattern Recognition and Artificial Intelligence Academic Press* 1976.
12. Davis SB, Mermelstein P: **Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentencesO**. *IEEE Transactions on Acoustics Speech and Signal Processing* 1980, **28**(4):357–366.
13. Rabiner L, et al.: **A tutorial on hidden Markov models and selected applications in speech recognition**. *Proceedings of the IEEE* 1989, **77**(2):257–286.
14. Potamianos G, Graf HP, Cosatto E: **An image transform approach for HMM based automatic lipreading**. In *IEEE International Conference on Image Processing*, Chicago, Il 1998:173–177.
15. Scanlon P, Ellis D, Reilly R: **Using mutual information to design class specific phone recognizers**. In *Proceedings of Eurospeech* 2003.
16. Patterson EK, Gurbuz S, Tufekci Z, Gowdy JN: **Moving-talker, speaker-independent feature study, and baseline results using the WAVE Multimodal speech corpus**. *Eurasip Journal on Applied Signal Processing* 2002, **2002**(11):1189–1201.
17. Sonka M, Hlavac V, Boyle R: **Image Processing, Analysis, and Machine Vision**. *International Thomson* 1999.
18. Lachenbruch P, Goldstein M: **Discriminant analysis**. *Biometrics* 1979, **35**:69–85.
19. Battiti R: **Using mutual information for selecting features in supervised neural net learning**. *IEEE transactions on neural networks* 1994, **5**(4):537–550.
20. Fleuret F: **Fast binary feature selection with conditional mutual information**. *The Journal of Machine Learning Research* 2004, **5**:1531–1555.
21. Peng H, Long F, Ding C: **Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy**. *IEEE transactions on pattern analysis and machine intelligence* 2005, **27**(8):1226–1238.
22. Gurban M, Thiran JP: **Information theoretic feature extraction for audio-visual speech recognition**. *IEEE Transactions on Signal Processing* 2009.
23. Adjoudani A, Benoit C: *Speechreading by humans and machines*, Springer 1996 :461–471.

24. Chen T: **Audiovisual speech processing**. *IEEE Signal Processing Magazine* 2001.
25. Neti C, Potamianos G, Luettin J, Matthews I, Glotin H, Vergyri D: **Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop**. In *Proc. Works. Signal Processing* 2001.
26. Potamianos G, Luettin J, Neti C: **Hierarchical discriminant features for audio-visual LVCSR**. In *ICASSP Proceedings* 2001.
27. Movellan J, Chadderdon G: **Channel separability in the audio-visual integration of speech: A Bayesian approach**. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES* 1996, **150**:473–488.
28. Massaro D, Stork D: **Speech recognition and sensory integration**. *American Scientist* 1998, **86**(3):236–244.
29. Kittler J, Hatef M, Duin R, Matas J: **On combining classifiers**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998, **20**(3):226–239.
30. Kirchhoff K, Bilmes J: **Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values**. *International Conference on Acoustics, Speech, and Signal Processing* 1999, :693–696.
31. Rabiner L, Juang B: **An introduction to Hidden Markov Models**. *IEEE ASSP Magazine* 1986, **3**:4–16.
32. Bishop C: *Neural Networks for Pattern Recognition*. Oxford University Press 1995.
33. Potamianos G, Neti C, Gravier G, Garg A, Senior A: **Recent Advances in the Automatic Recognition of Audio-Visual Speech**. *Proceedings of the IEEE* 2003, **91(9)**.
34. Furui S: **Cepstral analysis technique for automatic speaker verification**. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 2003, **29**(2):254–272.
35. Rabiner L, Juang BH: *Fundamentals of Speech Recognition*. Signal processing, Prentice Hall 1993.
36. Blanz V, Grother P, Phillips P, Vetter T: **Face recognition based on frontal views generated from non-frontal images**. In *IEEE Proceedings of Computer Vision and Pattern Recognition, Volume 2* 2005:454–461.
37. Gross R, Matthews I, Baker S: **Appearance-based face recognition and light-fields**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2004, **26**(4):449–465.
38. Blanz V, Vetter T: **Face recognition based on fitting a 3D morphable model**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2003, **25**(9):1063 – 1074.
39. Blanz V, Vetter T: **Face recognition based on fitting a 3D morphable model**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2003, :1063–1074.
40. Wai Lee M, Ranganath S: **Pose-invariant face recognition using a 3D deformable model**. *Pattern Recognition* 2003, **36**(8):1835–1846.
41. Vetter T: **Synthesis of novel views from a single face image**. *International Journal of Computer Vision* 1998, **28**(2):103–116.
42. Beymer D: **Face recognition under varying pose**. *IEEE Proceedings of Computer Vision and Pattern Recognition* 1994, :756 –761.
43. Chai X, Shan S, Chen X, Gao W: **Locally linear regression for pose-invariant face recognition**. *IEEE Transactions on Image Processing* 2007, **16**(7):1716–1725.
44. Potamianos G, Neti C, Gravier G, Garg A, Senior A: **Recent advances in the automatic recognition of audiovisual speech**. *Proceedings of the IEEE* 2003, **91**(9):1306 – 1326.
45. Summerfield Q: *Hearing by Eye: The Psychology of Lip-Reading*. Lawrence Erlbaum Associates 1987.
46. Bishop C, et al.: *Pattern recognition and machine learning*. Springer New York 2006.
47. Bellman R: **Adaptive control processes: a guided tour**. *Princeton University Press* 1961, **1**:2.
48. Young S, Evermann G, Kershaw D, Moore G, Odell J, Ollason D, Valtchev V, Woodland P: *The HTK book, Volume 2*. Citeseer 1997.
49. Jordan T, Thomas S: **Effects of horizontal viewing angle on visual and audiovisual speech recognition**. *Journal of Experimental Psychology* 2001, **27**(6):1386–1403.
50. Varga A, Steeneken H, Tomlinson M, Jones D: **The NOISEX-92 study on the effect of additive noise on automatic speech recognition**. *DRA Speech Research Unit, Malvern, England, Tech. Rep* 1992.
51. Bisani M, Ney H: **Bootstrap estimates for confidence intervals in ASR performance evaluation**. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* 2004.

## Figures

### Figure 1 - Sample figure title

A short description of the figure content should go here.

### Figure 2 - Sample figure title

Figure legend text.