MULTIPOSE AUDIO-VISUAL SPEECH RECOGNITION

Virginia Estellers, Jean-Philippe Thiran

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

ABSTRACT

In this paper we study the adaptation of visual and audio-visual speech recognition systems to non-ideal visual conditions. We focus on the effects of a changing pose of the speaker relative to the camera, a problem encountered in natural situations. To that purpose, we introduce a pose normalization technique and perform speech recognition from multiple views by generating virtual frontal views from non-frontal images. The proposed method is inspired by pose-invariant face recognition studies and relies on linear regression to find an approximate mapping between images from different poses. Lipreading experiments quantify the loss of performance related to pose changes and the proposed pose normalization techniques, while audio-visual results analyse how an audio-visual system should account for non-frontal poses in terms of the weight assigned to the visual modality in the audio-visual classifier.

1. INTRODUCTION

During the last few years a general framework for Audio Visual Automatic Speech Recognition (AV-ASR) has been developed [15], but a practical deployment has not yet taken place because systems lack robustness against non-ideal working conditions. Research has particularly neglected the variability of the visual modality subject to real scenarios, due in part to the lack of large corpora reproducing expected working conditions of the systems, i.e non-uniform lighting and non-frontal poses caused by natural movements of the speaker. Recently, however, works on meeting room scenarios and in-car vehicle systems made available more realistic data and enabled studies on genuine AV-ASR applications. The first studies of that kind [10, 14] applied directly the lipreading[1] systems developed for ideal visual conditions into a real scenario, obtaining poor lipreading results and failing to exploit the visual modality in the multimodal system. Those 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 models. It is necessary then to develop an effective framework for pose invariant lipreading instead of simply building feature extraction and classification blocks for each possible continuous pose. The same problem exists in the face recognition task and it then 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. Previous work on this topic is limited to Lucey et al [11–13], who applied linear regression (LR) to project visual speech features of complete profile images to a frontal viewpoint. In our work we introduce other projection techniques applied in face recognition to the lipreading system and compare their effects in different feature spaces: the images themselves, a smooth and compact representation of the images in the frequency domain or the final features used in the classifier. The main contibutions of our work are the extension of previous pose normalization techniques to intermediate poses at 30°, 60°and 90°of head rotation, the adaptation of other projection methods borrowed from face recognition and the evaluation of how non-frontal visual streams should be integrated into an audio-visual system in terms of the weight associated to the visual stream in the classifier.

The paper is organized as follows. In section 2 we present the existing and proposed techniques used to project non-frontal views to a frontal viewpoint. Section 3 explains how the pose-invariance is introduced in the lipreading system. Experimental results are reported in section 4 for visual and audio-visual systems and, finally, conclusions are drawn in section 5.

2. POSE-INVARIANT LIPREADING

The techniques proposed for pose-invariant face recognition can be classified into viewpoint transform and coefficient-based techniques [5]. The coefficient based approach estimates the face under all viewpoints given a single view by defining pose-invariant features [9]. On the other hand, viewpoint transform approaches use a face recognition system designed for the dominant view (frontal) and include a pre-processing step transforming the input images of undesired (non-frontal) poses to the desired view [5]. 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 benefit only a small fraction of time from pose-invariant features, while a system optimized for frontal views suits most of the time the working conditions.

Essentially there are two strategies to generate virtual frontal views from non-frontal poses: 3-Dimensional (3D) models [6] and learning-based methods [2, 18]. In the first case, a 3D morphable model of the face must be built from 2D images before virtual views from any viewpoint can be generated. It is computationally expensive and time consuming to match the input 2D image with the 3D model and, therefore, that technique is not aimed to most real-world applications. To overcome that issue, learning-based approaches learn how to estimate virtual views directly in the 2D domain, either via a 2D face model or from the images themselves. Several reasons favour last strategy in face recognition [5, 7] and justify the use of LR to project the images from lateral to frontal views in AV-ASR. 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 for the pose normalization step [15]. 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 [17]. At the same time, the proposed pose normalization involves transforms that can be quickly computed and allow real-time implementations. 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 has been applied to visual speech recognition

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

[1] Visual ASR or lipreading and Audio-Visual ASR systems share the same visual feature extraction and differ only on the statistical models used for classification, which are a combination of the audio and visual models used in single modality ASR. On the following, to avoid confusion, we refer to their common visual feature extraction and speech modelling blocks as lipreading system.

to project the visual speech features extracted from complete profile images to the frontal feature space [11–13]. We propose to extend these works in several ways. First, we analyse the performance of LR for different intermediate poses between fully frontal and fully profile views, we study the influence of applying LR on the images themselves or to different features spaces involved in the speech recognition system, see figure 1, and we finally propose a local version of LR and compare it to previous techniques.

Let's recall fisrt the basics of linear regression. 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

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

which measures the mean square error on the training dataset and might include a Tykhonov regularization term (weighted by parameter β) introducing additional smoothness properties and leading to a Ridge Regression [4]. 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, 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 3D nature of faces affect the different views and the linear mapping assumption no longer holds. To this end, [7] proposes the use of a piecewise linear function to approximate the non-linear mapping existing between images from different poses. The main idea of this 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. We refer to that technique as local LR (LLR) in opposition to the previous implementation of LR, which considered the images as a whole and is therefore designated as G LR (GLR).

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 $\{X_i,Y_i\}_{i=1\dots N}$ and an LR matrix W_i is computed for each pair of frontal and lateral patches. In the testing stage, the images are anew partitioned, each frontal patch is predicted with the corresponding matrix W_i and combined with other patches to obtain 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. 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).

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 lipreading system. A first set of features X_F,Y_F are designed to smooth the images and obtain a more compact and low-dimensional

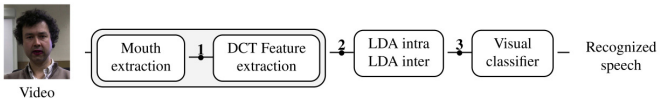


Figure 1: Lipreading system and feature spaces where pose-normalization is applied: 1 for images, 2 for DCT and 3 for LDA.

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. For more details about features used in lipreading, we refer the reader to [14].

The visual features X_F,Y_F are the first coefficients of the two-dimensional Discrete Cosine Transform (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 of the image X_I,Y_I and there is then an approximate linear mapping between the DCT coefficients of the frontal and lateral images. The linear relationship, however, no longer holds when we consider only a reduced set of DCT coefficients (first 140 coefficients out of 4096) and the transform W^F is only an approximation of the non-linear mapping existing between any pair of reduced DCT coefficients. 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* [1]. In that sense, the effect of the regularization parameter β is more important in the estimation of W^I than W^F. Here, the LLR technique provides a different meaning to the patches, namely frequency bands. 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. With no additional information, we choose an equal partition of the selected DCT coefficients to define the frontal and associated lateral patches in the LLR transform.

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 [15]. The transforms are usually based on Linear Discriminant Analysis (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

$$S_w = \sum_{i=1}^{C} p_i\Sigma_i \qquad S_b = \sum_{i=1}^{C} p_i\left(\mu - \mu_i\right)\left(\mu - \mu_i\right)^T, \qquad (2)$$

where p_i is the percentage of samples on the training set belonging to the class i, μ_i and Σ_i are the mean and covariance matrix for those samples and μ 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 if v is an eigenvector of R^y with eigenvalue λ_v, then $W^{-1}v$ is an eigenvector of R^x with the same eigenvalue. 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

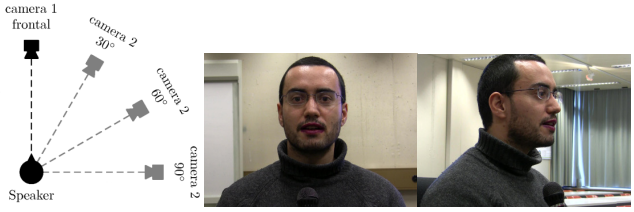
approximates the relation between X_F and Y_F. Second, two stages of LDA are needed to obtain X_L and Y_L from X_F and Y_F, a first intra-frame LDA and then an inter-frame LDA on concatenated adjacent vectors extracted from the intra-frame LDA. It is easy to prove that the linear relationship still holds if we consider now the transform on concatenated adjacent vectors of X_F and Y_F. We can then justify the use of LR to estimate the transform between the LDA feature spaces associated to different poses, which was missing up to the moment [11–13]. Observe 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 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 projection of LDA features might be able to exploit their specificity for the speech recognition task.

3. SPEECH RECOGNITION SYSTEM

Our lipreading system is composed of three blocks: the mouth detection and extraction, visual feature transformation and speech classification. For Audio-Visual ASR we have also the corresponding audio feature extraction, while the audio and visual fusion takes places in the classifier by means of a weighted multi-stream Hidden Markov Model (HMM). In our experiments we assume the pose to be known and introduce a pose normalization block. When the transformations are applied directly to the image space, the pose normalization takes place after the mouth extraction, whereas for the DCT or LDA features the transformation is introduced after the corresponding feature transform block.

For the audio modality, the system includes a state-of-the-art audio feature extraction block, where 13 Mel Frequency Cepstral Coefficients (MFCC) are extracted at an audio rate of 100 Hz with a 25 ms Hamming window. We append then their first and second time derivatives to include dynamic information and remove their means by Cepstral Mean Subtraction [8].

The first block of our visual system extracts images of the speaker's mouth from the original videos. It defines a mouth region-of-interest (ROI), which is then scaled in size, centred and rotated in order to obtain normalized mouth images for the different speakers, from which the visual speech features are afterwards extracted. Extraction of the mouth ROI constitutes part of the face tracking task and it is not a problem generally studied in lipreading. To that purpose, we work with sequences where the speaker wears blue lipstick and we can accurately track the mouth by color information in the hue domain. For each frame we estimate the position of the lips, the center and corners of the mouth, excluding outliers from the estimated positions over a sequence. Finally, a sequence of normalized 64x64 pixels ROIs centered on the mouth is extracted. On the following, we designate as F and L-sequences the obtained sequences for frontal and lateral poses at 30°, 60°and 90°of head rotation, which correspond to the previous x_I and y_I image vectors. Next, the system obtains a compact low-dimensional representation of the image x_F, y_F by extracting its first 140 DCT coefficients in zig-zag order. To normalize the features for different speakers, we remove their mean value over the sequence with the equivalent technique to the Cepstral Mean Subtraction. Finally, the 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 useful in speech recognition 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 size of the selected DCT coefficients, inter and intra-frame LDA parameters are chosen based on experiments with an evaluation dataset to optimize speech recognition.



(a) Camera positions (b) Speaker frontal pose (c) Speaker pose 60°

Figure 2: Schema of simultaneous recordings with different poses.

For the audio-visual classification, the system uses multi-stream HMMs [16] to combine both audio and visual streams. Weighted multi-stream HMMs are the natural extension of HMMs when two independent feature streams are defined as observations. They introduce stream weights λ_A, λ_V to controls their joint audio-visual observation model associated to the HMM state variable q by $p(o_A o_V|q) = p(o_A|q)^{\lambda_A} p(o_V|q)^{\lambda_V}$. To keep the same relationship between emission and transition probabilities as in single-stream HMMs, the weights are usually forced to sum-up to one and their value is proportional to the reliability associated to each stream for speech recognition. In order to quantify the loss of performance associated to each view, the same kind of visual classifiers is trained for each possible pose: frontal (abbreviated as F-class) and lateral at 30°, 60°and 90°of head rotation (L-class).

4. EXPERIMENTAL RESULTS

We perform connected speech recognition experiments under different speaker poses relative to the camera. To train and test the different methods we apply the multi-speaker paradigm (all speakers are on train and test set but with different sequences) with three fold cross-validation and give the results in terms of word accuracy. The same multi-speaker cross-validation is used to estimate the LR transforms for the different poses and features. We used the HTK tool-kit [19] 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 was used. The model parameters were initialized with the values estimated for independent audio and visual HMMs and posteriorly re-estimated jointly with four iterations of the expectation maximization algorithm. We considered the audio and visual weights fixed parameters of the system, restrict them to sum up to one and choose the weights leading to best speech recognition on an evaluation dataset.

4.1 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 (available on our webpage), consisting 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 2. 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 $3x3$ 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 1920x1080 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. Audio was recorded with a sample rate of 48000 Hz and 256 kbps for both micros, but only the clean audio signal obtained with the external microphone (Sony F-V120) was posteriorly used in the audio-visual experiments. We synchronized the videos from the audio signal because it offered better time resolution than a pairing of the video frames. For the two audio signals we computed the correlation of their normalized MFCC features within each manually segmented word, we then estimated a delay for each word and averaged over the whole sequence in order to obtain the delay between audio signals with a resolution of 10 milliseconds. 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.

4.2 Visual Speech Recognition

In a first set on experiments we paired the frontal and lateral sequences and test each sequence with the corresponding system, 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 column "Baseline" from table 1. As expected, speech recognition deteriorates with non-frontal speaker poses, which of course is more acute for 90°(10% of loss of performance) than for 60°(5% of loss of performance). It is interesting to note that there is no statistically significant[2] loss of performance between the frontal and the lateral sequences for 30°of head rotation. We also present 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 around 70% to 20%. Finally, we test the different pose normalization techniques with the L-sequences on the classifier trained and optimized for frontal sequences (" F-class, L-sequences" in table 1). 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.

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 and LDA features with $\beta = 5$ and 0 respectively. 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 3% 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 26% to 37%). The fact that the original images and the selected DCT coefficients present similar performance with different LR techniques and regularization parameter β 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 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. In terms of speech recognition it is in fact equivalent working on the high-dimensional image space with the local version of LR to applying the GLR on the reduced DCT space, essentially because any improvement on the virtual views obtained in the LLR projection of images is lost on their posterior projection to the reduced DCT space. Comparing the different LR techniques applied to the different spaces, we see that LLR performs better than GLR only for the original images, where the assumption of a piece-wise linear mapping can be related to images patches containing different parts of the mouth. For the DCT, however, the patches correspond to high and low-frequency components of the images and only a linear transform between the low-frequency components of the images can be justified, while that assumption does not hold for the high-frequency components associated to image details. In the case of intra-inter LDA features there is no interpretation of the patches defined on the LLR technique and the GLR and LLR techniques perform similarly.

Head	Baseline		F-class, L-sequences			
pose	F-seq	L-seq	L-seq	LR images	LR DCT	LR LDA
30 °	70.2	70.8	21.3	43.5	44.0	67.7
60 °	72.3	67.3	23.5	39.7	37.5	64.3
90 °	70.7	60.0	20.0	33.0	32.8	58.3

Table 1: Lipreading word accuracy (%) with different visual streams and classifiers. Comparison to Baseline quantifies the loss of associated to each pose-normalization technique and different levels of head rotation.

4.3 Audio-Visual Speech Recognition

We study how 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 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.

Table 2 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. Consequently, the differences in performance of the pose normalization methods are more acute with 0 dB than 7 dB 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

[2]Statistical analysis following [3] was performed for all the experiments but it is not included in the results due to space restrictions. However, when we mention that systems are equivalent or have the same performance, it is because the differences of performance across the different train/test sets and speakers are not statistically significant.

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 [3]. We also analyse the

Word accuracy	0 dB audio SNR			7 dB audio SNR		
Lipreading system	30°	60°	90°	30°	60°	90°
audio-only	36.5	37.0	36.2	64.2	67.5	67.3
F-seq F-class	71.5	78.3	70.3	78.2	84.6	79.7
L-seq L-class	72.0	71.7	64.3	78.5	79.7	74.8
LR images	44.7	46.5	37.3	56.0	56.1	46.8
LR DCT	44.9	46.5	37.5	57.2	56.3	48.7
LR LDA	68.3	72.0	65.8	77.0	77.7	75.6

Table 2: Word accuracy (%) for audio and audio-visual systems with different visual streams and classifiers.

value of the video weight λ_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 table 3, 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 λ_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°. In fact, there is a clear correlation between the values of the optimal visual weight and the stream's performance in lipreading experiments, as presented in figure 3. This figure shows that we can derive the optimal visual weight for each pose-normalization from its lipreading performance. Consequently, improvements in the visual lipreading system can be directly mapped to the corresponding audio-visual system by means of the weight associated to the visual stream.

λ_V	0 dB audio SNR			7 dB audio SNR		
Lipreading system	30°	60°	90°	30°	60°	90°
F-seq F-class	0.67	0.72	0.70	0.6	0.57	0.57
L-seq L-class	0.68	0.58	0.65	0.52	0.53	0.47
LR images	0.45	0.45	0.20	0.28	0.28	0.18
LR DCT	0.52	0.38	0.18	0.27	0.28	0.22
LR LDA	0.68	0.63	0.63	0.48	0.47	0.40

Table 3: Optimal video weight in the audio-visual systems with different visual streams and classifiers.

5. CONCLUSIONS

In this paper we presented a lipreading system able to recognize speech from different views of the speaker. We rely on pose normalization techniques used in face recognition to generate virtual frontal views of the speaker's mouth, or the corresponding speech features, from non-frontal images. Our experiments show that the pose normalization is more successful when applied directly to the LDA features used for speech recognition, while the more general feature spaces defined by the images themselves or their low-frequency representation suffer from misalignments of the training data or the estimation of the linear regression projection. We also study the integration of such a lipreading system into an audio-visual speech recognizer, quantifying the loss of performance related to pose changes and normalization techniques and how the weighting associated to the visual stream should account for it. The results obtained with the audio-visual system are coherent with the ones obtained in lipreading and, thus, any improvement obtained in the visual-only domain for pose normalization can be transferred into the audio-visual task by adapting the weight of the visual stream in the audio-visual classifier.

[3] 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.

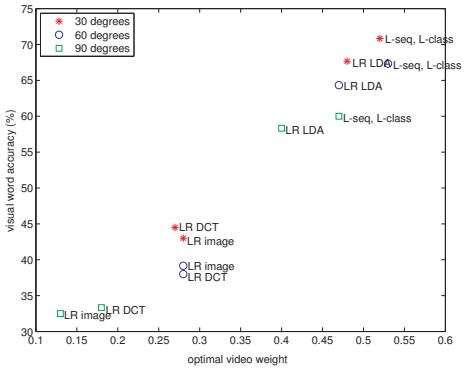


Figure 3: Scatter between optimal video weight and visual-only speech recognition performance for the different visual streams with a corrupted audio signal with 7 dB of SNR.

REFERENCES

[1] R. Bellman. Adaptive control processes: a guided tour. *Princeton University Press*, 1:2, 1961.

[2] D. Beymer. Face recognition under varying pose. pages 756 –761, 1994.

[3] M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *IEEE ICASSP Proceedings*, 2004.

[4] C. Bishop et al. *Pattern recognition and machine learning*. Springer, 2006.

[5] V. Blanz, P. Grother, P. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *IEEE CVPR Proceedings*, volume 2, pages 454–461, 2005.

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 1063–1074, 2003.

[7] X. Chai, S. Shan, X. Chen, and W. Gao. Locally linear regression for pose-invariant face recognition. *IEEE Trans. Image Processing*, 16(7):1716–1725, 2007.

[8] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoustics, Speech and Signal Processing*, 29(2):254–272, 2003.

[9] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(4):449–465, 2004.

[10] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, 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.

[11] P. Lucey, G. Potamianos, and S. Sridharan. A unified approach to multi-pose audio-visual ASR. In *IEEE Interspeech Proceedings*, pages 650–653, 2007.

[12] P. Lucey, G. Potamianos, and S. Sridharan. An Extended Pose-Invariant Lipreading System. In *International Workshop on Auditory-Visual Speech Processing*, 2007.

[13] P. Lucey, S. Sridharan, and D. Dean. Continuous Pose-Invariant Lipreading. In *IEEE Interspeech Proceedings*, 2008.

[14] G. Potamianos and C. Neti. Audio-visual speech recognition in challenging environments. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306 – 1326, 2003.

[16] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[17] Q. Summerfield. *Hearing by Eye: The Psychology of Lip-Reading*, chapter Some preliminaries to a comprehensive account of audio-visual speech perception. 1987.

[18] T. Vetter. Synthesis of novel views from a single face image. *International Journal of Computer Vision*, 28(2):103–116, 1998.

[19] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK book*, volume 2. 1997.