

Exploiting lower face symmetry in appearance-based automatic speechreading

Gerasimos Potamianos and Patricia Scanlon*

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

*Department of Electronic and Electrical Engineering, University College Dublin, Ireland

email: gpotam@us.ibm.com, patricias@ee.ucd.ie

Abstract

Appearance-based visual speech feature extraction is being widely used in the automatic speechreading and audio-visual speech recognition literature. In its most common application, the discrete cosine transform (DCT) is utilized to compress the image of the speaker's mouth region-of-interest (ROI), and the highest energy spatial frequency components are retained as visual features. Good generalization performance of the resulting system however requires robust ROI extraction and its consistent normalization, designed to compensate for speaker head-pose and other data variations. In general, one expects that the ROI - if correctly normalized - will be nearly laterally symmetric, due to the approximate symmetry of human faces. We thus argue that forcing lateral ROI symmetry can be beneficial to automatic speechreading, providing a mechanism to compensate for small face and mouth tracking errors, which would otherwise result to incorrect ROI normalization. In this paper, we propose to achieve such ROI symmetry indirectly, by considering the spatial frequency domain and exploiting the DCT properties. In particular, we propose to remove the odd frequency DCT components from the selected visual feature vector. We experimentally demonstrate that, in general, this approach does not hurt speechreading performance, while it reduces computation, since it results to less DCT features. In addition, for the same number of features, as in traditional DCT coefficient selection, the method results in significant speechreading improvements. For the connected-digit automatic speechreading experiments considered, and for low feature dimensionalities, such can reach up to 12% relative reduction in word error rate.

1. Introduction

Visual speech information, extracted from the speaker's mouth region, has been repeatedly demonstrated to improve performance and noise robustness of *automatic speech recognition* (ASR) [1]–[6]. Critical however to the performance of the resulting *audio-visual ASR* system is the choice of visual features that contain sufficient information about the uttered speech.

A widely used approach to extracting visual features constitute the so-called appearance-based techniques [2]. In their most typical form, these methods first require the extraction of a *region-of-interest* (ROI), usually containing the speaker's mouth and possibly its neighboring lower face (see also Fig. 1). In contrast to shape-based techniques [2], the appearance-based approach considers all pixels within the ROI as informative



Figure 1: Typical steps for extracting the visual speech region-of-interest (ROI), applied to sample video frames from the “studio” (upper) and “office” (lower) audio-visual datasets considered in this paper (see also Table 1). The following are depicted at each row, left-to-right: Original frame with detected facial feature points superimposed; face-area enhanced frame; normalized extracted ROI.

of the utterance, and seeks linear transformations of their values in order to represent the speech information in a compact, low-dimensional feature vector. Popular image transforms employed in this framework are *principal component analysis* (PCA) [3] and the *discrete cosine transform* (DCT) [4]–[6]. The latter is very widely used because, unlike PCA, is amenable to fast computations and avoids expensive training.

In the traditional use of DCT for automatic speechreading and audio-visual ASR, a small number of DCT coefficients are used as visual speech features, selected on basis of highest energy, such energy values computed over a set of training images [5]. Alternative selection schemes using DCT coefficient variance have also been considered, but have been reported inferior to the energy-based method [6]. Of course, the energy-based criterion is more appropriate for image compression, and is not designed to provide optimal discrimination between the speech classes of interest. This fact has motivated recent work by Scanlon et al. [7], that investigated the mutual information criterion for DCT feature selection instead. Interestingly, this research demonstrated that particular DCT coefficients located at the even columns of the two-dimensional spatial frequency lattice exhibit significantly higher mutual information than their neighbors located at the odd columns, although the latter have large average energy and are therefore selected by the energy criterion. This work also showed that low-dimensional feature vectors obtained by means of mutual information significantly

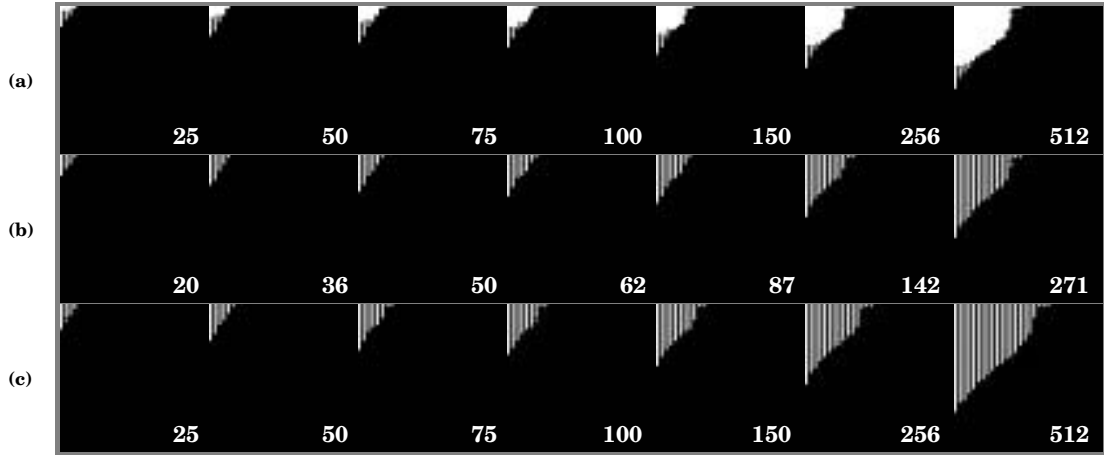


Figure 2: Examples of templates considered in this paper for DCT-based feature extraction. White elements denote locations of the retained DCT coefficients on the two-dimensional 64×64 frequency domain lattice. Three sets of templates are depicted, nested according to decreasing energy: (a): Baseline templates, using both odd and even locations; (b): Symmetric subsets of (a), with the odd components discarded; (c): Symmetric templates with equal number of elements as in (a). The number of elements of each template are also shown.

outperform the ones selected on basis of energy. However, no gain is achieved for feature dimensions higher than thirty. In addition, the method has the drawback that estimating mutual information of continuous-valued DCT coefficients is computationally expensive, and in any case significantly more involved than computing plain energy.

These observations have motivated us to consider a simple modification to the energy based selection criterion: namely, to discard the coefficients located at the odd columns of the two-dimensional spatial frequency lattice, and perform the selection on basis of decreasing energy over the pool of the remaining coefficients, located at the even columns. It is hoped that this modified criterion will result in DCT selection templates that - to an extent - mimic the mutual information based ones, while retaining the computational simplicity of the energy based selection approach. Investigating this modified technique for selecting visual speech features is the subject of this paper.

In more detail, in Section 2, we provide the motivation for discarding the odd components on basis of face symmetry. We argue that ROI symmetry is desirable and show that the DCT odd components of such ROIs are zero. In Section 3, we experimentally demonstrate the benefits of the modified selection scheme. We consider two multi-speaker connected-digit audio-visual databases, and we provide a number of visual-only recognition results for various DCT feature dimensionalities and feature post-processing schemes. Finally, Section 4 summarizes the paper.

2. ROI Symmetry and DCT Coefficient Selection

As discussed in the Introduction, most DCT-based automatic speechreading systems apply the transform on a two-dimensional ROI that contains the speaker’s mouth and possibly neighboring parts of the lower face, such as the cheeks and jaw

(see for example, Fig. 1). In general, human faces exhibit near lateral symmetry, one therefore expects that the extracted ROI should be approximately symmetric. Lateral symmetry could in fact be beneficial to speech classification, as it can be considered a desirable property of properly normalized ROIs, compensating for certain head pose and lighting variations. However, accurate ROI normalization is difficult, since for example tracking the face and facial landmarks is prone to error. One could therefore consider a ROI post-processing step in the pixel space, that could force the desired symmetry. In this work, we propose to achieve this symmetry operating in the spatial frequency domain instead, taking advantage of DCT properties. In the following, we discuss how to achieve this, and present in more detail the proposed DCT coefficient selection algorithm.

2.1. DCT of Symmetric Images

Let us first consider a one-dimensional discrete signal f_n , $n = 0, 1, \dots, N - 1$, of length N being a power of two. The one-dimensional discrete cosine transform of f , as computed in our automatic speechreading system (following [8]), is

$$F_k = \sum_{n=0}^{N-1} f_n \cos \frac{\pi k(2n+1)}{2N}, \text{ for } k = 0, 1, \dots, N.$$

It is not hard to see, that if the signal is symmetric around its “mid-point”, $(N - 1)/2$, i.e., $f_n = f_{N-n-1}$, for $n = 0, 1, \dots, N/2 - 1$, then

$$F_k = \sum_{n=0}^{N/2-1} f_n \left[\cos \frac{\pi k(2n+1)}{2N} + \cos \left(\pi k - \frac{\pi k(2n+1)}{2N} \right) \right]$$

$$= \begin{cases} 2 \sum_{n=0}^{N/2-1} f_n \cos \frac{\pi k(2n+1)}{2N}, & \text{if } k \bmod 2 = 0; \\ 0, & \text{if } k \bmod 2 = 1, \end{cases}$$

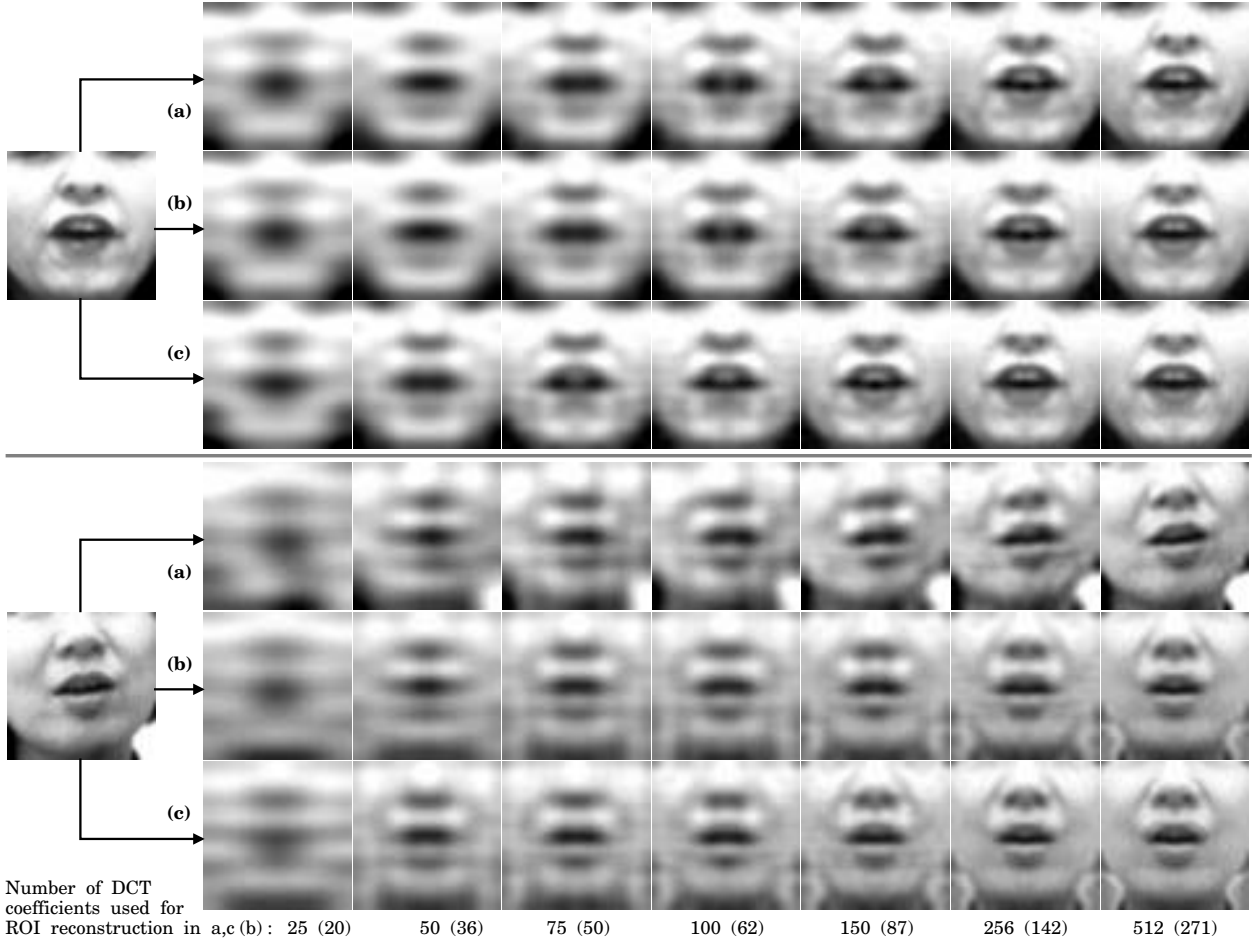


Figure 3: Examples of ROI reconstruction using the DCT templates of Fig. 2. Two ROI cases are depicted: In the upper part, the ROI is nearly laterally symmetric; in the lower part, the ROI is highly unsymmetric, partially due to poor normalization. Notice that the DCT templates of Fig. 2(b,c) provide a laterally symmetric ROI reconstruction. In general, the templates of Fig. 2(c) produce a clearer reconstruction than the ones of Fig. 2(a), using the same number of DCT coefficients. The templates of Fig. 2(b) provide similar quality of reconstruction as the baseline (a), but with fewer coefficients.

since

$$\cos \frac{\pi k(2n+1)}{2N} = (-1)^k \cos\left(\pi k - \frac{\pi k(2n+1)}{2N}\right). \quad (1)$$

Therefore, the DCT odd frequency components of a symmetric one-dimensional signal are all zero.

Similarly, if $F_k = 0$, for $k = 1, 3, \dots, N-2$ (assuming again that N is a power of two), then the inverse DCT, given by

$$f_n = \frac{1}{N}F_o + \frac{2}{N} \sum_{k=1}^{N-1} F_k \cos \frac{\pi k(2n+1)}{2N},$$

for $k = 0, 1, \dots, N$, becomes

$$f_n = \frac{1}{N}F_o + \frac{2}{N} \sum_{k=1}^{N/2-1} F_{2k} \cos \frac{\pi k(2n+1)}{N},$$

and therefore (see also (1))

$$f_n - f_{N-n-1} = \frac{2}{N} \sum_{k=1}^{N/2-1} F_{2k} \left[\cos \frac{\pi k(2n+1)}{N} - \cos\left(2\pi k - \frac{\pi k(2n+1)}{N}\right) \right] = 0,$$

for $n = 0, 1, \dots, N/2 - 1$. Hence, zero odd frequency DCT components imply a symmetric original signal.

The above results readily generalize to the two-dimensional DCT, which is applied to the ROI for speechreading. The transform is separable, i.e., the result can be obtained by one-dimensional DCTs, first applied to the ROI rows, followed by an application to the columns of the result. Clearly therefore, if the ROI is laterally symmetric, the odd columns at the intermediate step will be all zeros, and therefore the one-dimensional DCT applied to them will keep them unchanged. For the inverse transform, the same argument carries through, hence setting the spatial frequency DCT coefficients located at the odd columns equal to zero, results to a laterally symmetric reconstructed ROI.

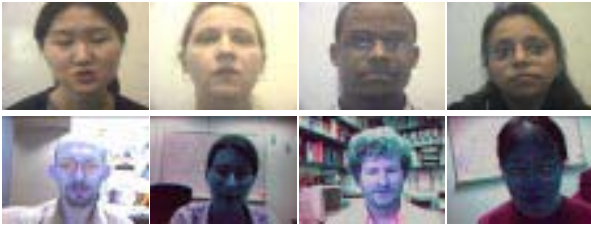


Figure 4: Example video frames from the “studio” (upper) and “office” (lower) audio-visual corpora (see also Table 1).

2.2. DCT Coefficient Energy-Based Selection Strategies

In the traditional use of DCT-based features for automatic speechreading and audio-visual ASR, given the mouth ROI, a small number of DCT coefficients are selected as visual features in order to produce a compact feature vector appropriate for speech classifier training. The selected locations on the two-dimensional spatial frequency lattice - typically consisting of 64×64 elements [5] - are chosen to correspond to the highest DCT coefficient energies, as estimated over a training set of video ROIs. In practice, the locations are sorted by decreasing energy, with the first n providing the n -element DCT selection template.

In the *baseline* approach widely followed in the literature, these energy values are computed over the entire frequency lattice. It turns out that odd column components have significant energies, since in practice the ROIs are never symmetric. This is due to the fact that perfect symmetry in nature is rare, and of course due to head-pose and lighting variations, that in general are extremely difficult to accurately compensate for (see also Fig. 1). As a result, odd column frequency components are present in the DCT selection templates, even for low-dimensional feature vectors (see also Fig. 2(a)).

In this paper, we investigate whether forcing ROI lateral symmetry is beneficial for automatic speechreading. Given the derivations above, this is equivalent to discarding the DCT coefficients located at the odd columns. This results to DCT templates that will be referred to as “*symmetric*”, and are derived by considering the energies at the even only columns of the two-dimensional frequency lattice. One can compare the performance of the resulting speechreading system to the baseline approach in two ways: First, by considering symmetric templates that constitute subsets of the baseline ones, obtained by plainly discarding their odd column components, as depicted in Fig. 2(b). If odd components are unnecessary, speechreading performance should not degrade when using the lower-dimensional symmetric templates, and one could argue that the algorithm derives a more compact visual speech feature representation, therefore reducing computations. Second, one can compare symmetric and baseline templates of the same size (see also Fig. 2(c)). If ROI symmetry is beneficial, it would be expected that symmetric template systems should outperform the baseline. Experiments along these two lines are presented in detail in Section 3.

To help visualize the differences between the baseline and symmetric DCT templates, Fig. 3 depicts ROI reconstruction

| DB Set | “Studio” | | | “Office” | | |
|--------|----------|--------|------|----------|--------|------|
| | Subj. | Utter. | Dur. | Subj. | Utter. | Dur. |
| train | 50 | 5403 | 7:53 | 101 | 4591 | 6:07 |
| test | 50 | 623 | 0:55 | 101 | 537 | 0:43 |

Table 1: Details of the two audio-visual databases (DB: “studio” and “office”) used for training and testing in the experiments reported in this paper: Number of subjects (identical for training and testing), number of utterances, and total duration (in hours) are depicted for each set.

using the templates of Fig. 2. There, two ROI cases are considered: In the upper part, the ROI is nearly symmetric with respect to its middle column. ROI reconstruction using baseline templates (a), or their subset symmetric templates (b), look visually similar. However, in the lower part, where the ROI contains a rotated mouth, non-uniform lighting and some background, the reconstructed ROI on basis of templates (b) becomes symmetric. In both cases, reconstruction using symmetric templates (c) (i.e., with the same number of elements as the baseline) look significantly more detailed than using the baseline.

3. Experiments

In order to study the benefits of the proposed approach, we conduct a number of visual-only recognition experiments on two small-vocabulary (connected digits) audio-visual databases. Before reporting our results, we provide a brief description of the automatic speechreading system components and of the two databases.

3.1. The Automatic Speechreading System

Given full-face video of the speaker’s face, our system first utilizes a statistical face and facial feature detector to estimate the location of the face and of landmark facial features. Given these, a greyscale 64×64 -pixel ROI is extracted, normalized for lighting, as well as for head size and orientation (see also Fig. 1). A two-dimensional DCT is then applied to the ROI, resulting in a 4096-dimensional vector of transform coefficients. Coefficient selection is then performed using the schemes discussed in the previous section. The selected coefficients are subsequently mean-normalized and interpolated to 100 Hz, providing what we refer to as “static” visual features.

To improve recognition performance, dynamic speech information needs to be also included into the visual feature vector. In our typical automatic speechreading system [5], this is achieved by a cascade of a linear discriminant analysis (LDA) projection, followed by a maximum likelihood linear transform (MLLT) rotation, that is applied to the concatenation of neighboring static visual feature vectors. This is referred to as the *inter-frame* LDA/MLLT. In the experiments considered in this paper, the cascade is assumed to result to final 41-dimensional visual features, when applied over a window of 11 consecutive static vectors. Notice though, that when the static feature dimension increases, the LDA matrix becomes quite large; alternatively, one can first apply an LDA/MLLT cascade to the static feature vector, i.e., an *inter-frame* LDA/MLLT. In this pa-

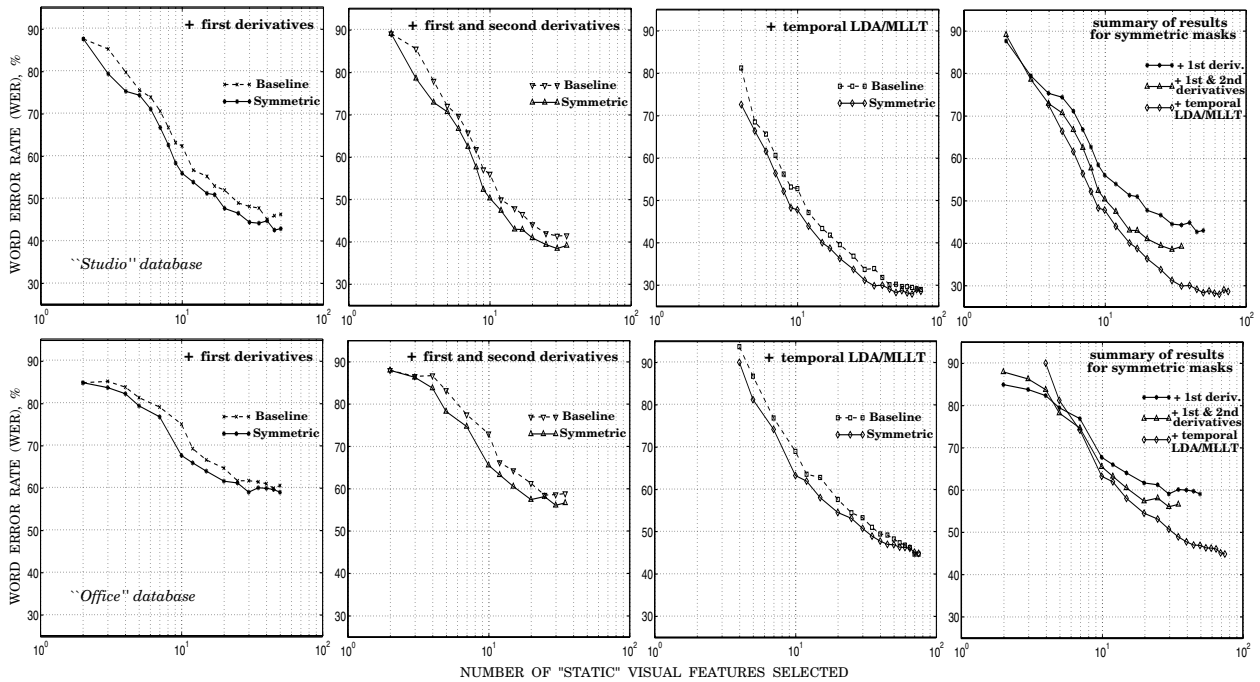


Figure 5: Visual-only word error rate (WER), %, on the test sets of the “studio” (upper plots), and “office” (lower) connected digit databases, using the baseline and symmetric DCT templates, depicted against the number of selected DCT coefficients (“static” DCT features). The selected features are fed into the automatic speechreading system after post-processing - in order to capture speech dynamics - by (depicted left-to-right): appending their first derivatives; appending both their first and second derivatives; inter-frame LDA/MLLT over 11 neighboring feature frames, producing 41-dimensional features. For easier comparisons, the WERs using the symmetric DCT templates are summarized at the right-most plots.

per, its output is assumed to be of dimension 30 [5]. In addition to these two techniques, the traditional approach of incorporating dynamic speech information is also considered, namely appending to the static visual feature vector its first- and possibly second-order derivatives.

The dynamic visual features are then fed into a *hidden Markov model* (HMM)-based recognizer. In our system, the HMMs are 3-state left-to-right phone models, consisting of context-dependent sub-phonetic states, with their parameters estimated by the expectation-maximization algorithm on available training data (see also Table 1). For the particular recognition task at hand (connected-digit strings), the HMMs correspond to 22 phones, 66 sub-phonetic states (3 states per phone), and 159 context-dependent states with 3.2k Gaussian mixture components. Decoding is performed using a stack decoder with an 11-word vocabulary (digits 0-9 and “oh”) and free grammar.

3.2. The Audio-Visual Database

We consider two corpora in our experiments containing audio-visual data of multiple speakers uttering connected digit sequences, recorded in two environments. The first set contains high-quality frontal full-face video of 50 subjects and is recorded in a quiet “studio”-like environment using a high-quality camera, uniform background and lighting, and relatively stable frontal subject pose. The video is MPEG2-encoded at a resolution of 704×480 pixels, and at 30 Hz.

The second set contains lower-quality video of 101 sub-

jects, collected using an inexpensive web-cam that is connected to a laptop-based data collection system through a USB-2.0 interface. Raw RGB data at 30 Hz but at a 320×240 pixel resolution are obtained, and subsequently MPEG1-encoded for processing. In contrast to the “studio” set, the data are collected under varying lighting and cluttered backgrounds, typically in the subjects’ offices. The set is therefore referred to as the “office” database, and it obviously represents a significantly more challenging visual domain than the previous set.

The experimental framework on the two corpora is summarized in Table 1, and assumes a multi-speaker training/testing paradigm. Example video frames from the two datasets are depicted in Fig. 4 (see also Fig. 1).

3.3. Results and Discussion

We now proceed to compare automatic speechreading performance based on the baseline and symmetric DCT coefficient selection mechanisms. As discussed above, we perform these comparisons using up to four methods of visual feature post-processing in order to incorporate speech dynamics into the speechreading system, and over two databases providing both visually clean and challenging data. The experimental results are summarized in Fig. 5 and in Table 2.

In Fig. 5 we compare visual-only word error rate (WER) on the “studio” and “office” corpora test sets between the baseline and symmetric DCT templates of the same size (number of static visual features), using three methods of including speech

| n | dct | Δ^1 | Δ^2 | Λ^1 | Λ^2 | n | dct | Δ^1 | Δ^2 | Λ^1 | Λ^2 |
|-----|-----|------------|------------|-------------|-------------|-----|-----|------------|------------|-------------|-------------|
| 10 | (a) | 62.5 | 56.1 | 52.9 | — | 50 | (a) | 46.4 | — | 30.3 | 30.7 |
| 8 | (b) | 62.7 | 57.7 | 52.2 | — | 36 | (b) | 44.8 | 38.9 | 30.1 | 30.3 |
| 10 | (c) | 56.1 | 50.4 | 47.9 | — | 50 | (c) | 43.0 | — | 28.4 | 28.2 |
| 20 | (a) | 52.1 | 44.2 | 39.7 | — | 75 | (a) | — | — | 29.1 | 28.8 |
| 16 | (b) | 51.5 | 43.5 | 39.8 | — | 50 | (b) | 43.0 | — | 28.4 | 28.2 |
| 20 | (c) | 47.8 | 41.1 | 36.4 | — | 75 | (c) | — | — | 28.7 | 28.3 |
| 35 | (a) | 47.9 | 41.6 | 34.0 | 34.1 | 100 | (a) | — | — | — | 28.6 |
| 25 | (b) | 46.7 | 39.5 | 33.8 | — | 62 | (b) | — | — | 28.5 | 28.6 |
| 35 | (c) | 44.3 | 39.3 | 30.0 | 29.9 | 100 | (c) | — | — | — | 28.2 |

Table 2: Visual-only word error rate, %, on the “studio” database test set using different-size templates for selecting n DCT coefficients of the ROI (see also Fig. 2): (a): baseline templates with both odd and even components; (b): symmetric templates, subsets of the ones of (a), with the odd frequency elements removed; (c): symmetric templates with equal number of DCT coefficients as the baseline. The resulting features are fed into the automatic speechreading system after post-processing (in order to capture speech dynamics) by: (Δ^1): appending their first derivatives; (Δ^2): appending both their first and second derivatives; (Λ^1): inter-frame LDA/MLLT over 11 neighboring feature frames; (Λ^2): both intra- and inter-frame LDA/MLLT. Missing entries are due to n being outside the range for the particular post-processing technique (see also Fig. 5).

dynamics into the recognition. As it is clear, the proposed technique that employs symmetric DCT templates consistently outperforms the baselines in all cases, especially for low and moderate sized feature vectors. The improvements reach up to a 12% relative WER reduction, for example for $n = 35$, in the “studio” dataset, when using inter-frame LDA/MLLT (34.01% \rightarrow 29.96%). Note that the baseline and symmetric templates are identical in their first two feature locations. However, the third location of the baseline template corresponds to an odd frequency component, thus the two templates start differing for values of $n \geq 3$. Notice also that the particulars of each feature post-processing technique pose limitations on the range of n . In general though, the inter-frame LDA/MLLT outperforms the derivative-based approaches, as is also clear from the summary plots to the right. Finally, performance on the “office” data lags significantly behind the one on the “studio” environment.

Table 2 depicts some of the WER numbers in Fig. 5 for particular templates in the case of the “studio” test set. All four feature post-processing techniques are now depicted. Notice, that for the particular test set size, a difference of about 1.6% in WER is significant at the 5% level. By comparing the entries in the (a) and (b) rows of the table, it becomes clear that, in general, discarding the odd column frequency components from the baseline templates, in order to obtain symmetric ones, does not hurt performance in a statistically significant way (with one exception for $n = 10 \rightarrow 8$); in contrast, it often helps. And of course, as in Fig. 5, when comparing rows (a) and (c), i.e., baseline and symmetric templates with the same number of features, the latter are consistently superior.

4. Summary

In this paper, we proposed to exploit lateral symmetry of the mouth region of interest in order to provide more compact visual feature representation and improve the resulting speechreading performance. Instead of pursuing symmetry in the image domain, we proposed to force such in the spatial frequency domain instead, taking advantage of DCT properties, by discarding DCT coefficients located at the odd frequency columns and selecting among the remaining DCT features on basis of energy. We compared this approach to a baseline, energy-based DCT feature selection over the entire spatial frequency lattice (i.e., including the odd columns), and we experimentally demonstrated that the proposed technique provides significantly better speechreading performance when the same number of visual features are used in both. The improvement was more pronounced for low to medium feature dimensions, where it reached an up to a 12% relative reduction in word error rate. In addition, in most cases, no degradation in performance was observed, when using lower dimensional feature vectors by just discarding the odd column frequency components of the baseline feature vector. The results were shown to hold for a number of visual feature post-processing techniques, as well as for both visually clean and challenging data.

5. References

- [1] E.D. Petajan, “Automatic lipreading to enhance speech recognition,” *Proc. Global Telecomm. Conf.*, pp. 265–272, 1984.
- [2] M.E. Hennecke, D.G. Stork, and K.V. Prasad, “Visionary speech: Looking ahead to practical speechreading systems,” in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke, Eds. Berlin: Springer, pp. 331–349, 1996.
- [3] C. Bregler and Y. Konig, “‘Eigenlips’ for robust speech recognition,” *Proc. Int. Conf. on Acoustics, Speech, and Signal Process.*, pp. 669–672, 1994.
- [4] P. Duchnowski, U. Meier, and A. Waibel, “See me, hear me: Integrating automatic speech recognition and lipreading,” *Proc. Int. Conf. Spoken Lang. Process.*, pp. 547–550, 1994.
- [5] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, “Recent advances in the automatic recognition of audiovisual speech,” *Proc. IEEE*, 91(9): 1306–1326, 2003.
- [6] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, “DCT-based video features for audiovisual speech recognition,” *Proc. Int. Conf. Spoken Lang. Process.*, pp. 1925–1928, 2002.
- [7] P. Scanlon, G. Potamianos, V. Libal, and S.M. Chu, “Mutual information based visual feature selection for lipreading,” *Proc. Int. Conf. Spoken Language Processing*, pp. 857–860, 2004.
- [8] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge Univ. Press, Cambridge, 1988.