DCT-BASED VIDEO FEATURES FOR AUDIO-VISUAL SPEECH RECOGNITION

Martin Heckmann[◇•], *Kristian Kroschel*[•]

Christophe Savariaux[◇], *Frédéric Berthommier*[◇]

[•]Universität Karlsruhe
Institut für Nachrichtentechnik
Kaiserstraß e 12, 76128 Karlsruhe, Germany
{heckmann, kroschel}@int.uni-karlsruhe.de

[◇]Institut National Polytechnique de Grenoble
Institut de la Commuinication Parlée
46, Av. Félix Viallet, 38031 Grenoble, France
{savario, bertho}@icp.inpg.fr

ABSTRACT

Encouraged by the good performance of the DCT in audio-visual speech recognition [1], we investigate how the selection of the DCT features influences the recognition scores in a hybrid ANN/HMM audio-visual speech recognition system on a continuous word recognition task with a vocabulary of 30 numbers. Three sets of features, based on the mean energy, the variance and the variance relative to the mean value, were chosen. The performance of these features is evaluated in a video only and an audio-visual recognition scenario with varying *Signal to Noise Ratios (SNR)*. The audio-visual tests are performed with 5 types of additional noise at 12 SNR values each. Furthermore the results of the DCT based recognition are compared to those obtained via chroma-keyed geometric lip features [2]. In order to achieve this comparison, a second audio-visual database without chroma-key has been recorded. This database has similar content but a different speaker.

1. INTRODUCTION

The importance of the lips movement in human speech perception, especially in noisy conditions, is well known. Therefore many researchers also take advantage of the visual information in *Automatic Speech Recognition (ASR)* systems. This article focuses on the extraction of the visual information for the recognition process and its compatibility with a HMM/ANN recognition system. The presented visual feature extraction is based on a *Discrete Cosine Transform (DCT)*. We evaluate different strategies to chose those DCT coefficients which are best suited for the recognition process. We analyze their efficiency in a video only and in different audio-visual recognition scenarios. Furthermore we try to draw a comparison between these DCT features and geometric lip features with different noise types at different SNR levels. This extends the study in [1] performed on video alone for a set of pixel based vs. geometric methods. Our comparison is based on two similar datasets. For the first dataset we use the DCT to extract the relevant visual features and in the second the lips of the speaker are colored with blue ink which allows to precisely extract the geometric lip parameters.

2. RECOGNITION SYSTEM

The recognition tests aiming to assess the performance of the DCT features for audio-visual speech recognition are carried out with an ANN/HMM hybrid model for continuous number recognition. Identification of the phonemes is performed independently for the audio and the video path (compare Fig. 1) and thus follows a SI or multi-stream approach [3]. The ANNs are trained to produce estimates of the a-posteriori probabilities $P(H_i|\mathbf{x}_A)$ and $P(H_i|\mathbf{x}_V)$ for the occurrence of the phoneme H_i when the acoustic feature vector $\mathbf{x}_A$ and the visual feature vector $\mathbf{x}_V$ are observed, respectively.

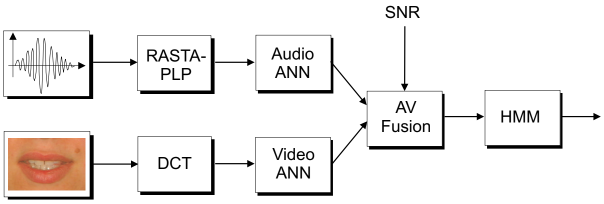

Fig. 1. Separate Identification (SI) audio-visual speech recognition system

Fusion of the estimated audio and video a-posteriori probabilities follows a Geometric Weighting [4]:

$$\hat{P}_{GW}(H_i|\mathbf{x}_A,\mathbf{x}_V) = \frac{\hat{P}^{\alpha}(H_i|\mathbf{x}_A)\hat{P}^{\beta}(H_i|\mathbf{x}_V)}{\hat{P}^{\alpha+\beta-1}(H_i)} \cdot \varepsilon(\alpha,\beta) \tag{1}$$

The weighting parameters α and β both depend on a third parameter c according to:

This work was partly funded by the EC program SPHEAR and is a part of the project RESPITE

$$\alpha = \begin{cases} 0 & c < -1 \\ 1+c & -1 \leq c \leq 0 \\ 1 & c > 0 \end{cases}, \ \beta = \begin{cases} 1 & c < 0 \\ 1-c & 0 \leq c \leq 1 \\ 0 & c > 1 \end{cases} \tag{2}$$

The parameter c controls the weighting of audio and video data. When $c \simeq -1$ only the video signal contributes to the recognition, whereas for $c \simeq 1$ the recognition relies completely on the audio signal. $\epsilon(\alpha, \beta)$ is a normalization parameter independent on the actual phoneme. Implementation of the system was carried out with the tool STRUT from TCTS lab Mons, Belgium [5].

To train the ANNs and to perform the recognition tests we used a single-speaker audio-visual database recorded at the Institut de la Communication Parl´ee (ICP) in Grenoble, France. As in our previous study, this database is a repetition of the same set of utterances selected from Numbers95 [6]. These utterances were spoken by a native English-speaking female subject. Only half of this database, corresponding to 708 utterances, is currently formatted. Each of these 708 utterance consists of several continuously uttered numbers yielding to a total of 2508 words. The database was divided in a set of 516 utterances for training and a set of 192 utterances for testing. To facilitate the tracking of the mouth region the movements of the speakers head were restricted during the recordings via a helmet and a marker was positioned in the speakers face in a location not affected by the articulatory movements (see also Fig. 2a). A lamp positioned in front of the speaker ensured constant lightning conditions and high contrast images.

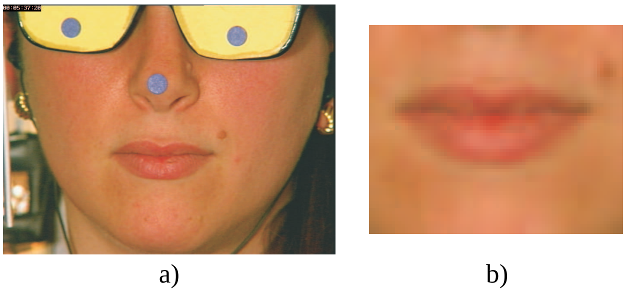
a) b)

Fig. 2. The blue markers on the glasses visible in the full size image a) were used to track the mouth region shown in b)

The recordings were made on BETACAM video with 50 half-frames of size 768 × 288 pixels per second. Instead of combining two half-frames to a full-frame we preferred to do without the additional spatial resolution and keep a higher temporal resolution. Full-frames were generated via an interpolation of the missing lines in each half-frame. After localization of the mouth region based on the markers positioned in the speakers face, the corresponding region was extracted and the images were down-sampled by a factor 4. This yields images of 78 × 64 pixels at 50 frames per second of the *Region Of Interest (ROI)* (compare Fig. 2b). Tracking of the ROI was performed via a correlation. The localization error in the final image is approximately 1 pixel.

3. VIDEO FEATURE EXTRACTION

3.1. The DCT

The extraction of the video features is performed with the *Discrete Cosine Transform (DCT)* [7]. The reasons for the widespread use of the DCT as well in image compression [8] as feature extraction [1] are the high compaction of the energy of the input signal onto a few DCT coefficients and the availability of a fast implementation of the transform, similar to the *Fast Fourier Transform (FFT)* [7]. Since the DCT is not shift invariant the performance depends on a precise tracking of the ROI.

3.2. Feature Types

As features for our recognition experiments we selected DCT coefficients following three different strategies:

energy features: the L features with the highest energy

variance features: the L features with the highest variance

relative variance features: the L features with the highest variance after normalization to their mean value

The number of features L extracted from each image frame was varied between 20 and 100. The necessary mean values and variances were calculated over the complete training set. Synchronization between the audio and video stream was obtained via an interpolation of the DCT coefficients to the audio feature rate.

4. RECOGNITION RESULTS

We performed two different types of recognition tests. In the first test we used only the video information and compared the recognition results of the different feature types with varying feature size. For the second test we combined the audio and video information and performed tests with varying SNR in the audio stream.

4.1. Video Only Tests

In Tab. 1 the WERs obtained by the different feature types with varying feature size are displayed. All tests were performed with a hidden layer of 5000 neurons which gave also for the larger feature sizes the best results. A time window of 13 frames set up by the current frame and the 6 preceding and succeeding frames was presented to the ANN. Additionally to the pure DCT coefficients also their delta and delta-delta values were used. As can be seen from Tab. 1 the

	20	30	40	60	100	average
Energy	32%	28%	32%	33%	32%	31.6%
Variance	35%	31%	32%	35%	30%	32.6%
Relative Var.	44%	40%	42%	40%	40%	41.2%

Table 1. Video recognition rates in percent WER with different feature types using between 20 and 100 features per frame. In the last column the average WER for a chosen feature type is given

results show only small variations with the feature size. No increase in performance with an augmentation of the number of features is visible. Taking the good results obtained with 30 features and the increase of computation time with an increase of the feature size into account, the choice of 30 features seems to be preferable and is used for subsequent tests. When looking at the different feature types the performance of the features based on the relative variance is clearly inferior to the other two feature types. The average WER over all tested feature sizes indicates a slightly better performance of the energy features. This better performance is also visible at the chosen size of 30 features. Overall this test clearly shows that the DCT coefficients are well suited for use in an ANN/HMM recognition system.

4.2. Audio-Visual Tests

In order to investigate the potential of the DCT based video features to take advantage of the complementarity of the audio and video data and to test their compliance with the Geometric Fusion presented in Sec. 2 we also performed audio-visual tests. For these tests 5 different noise types at 12 SNR values each were added to the audio signal and the fusion parameter c was adapted to each noise scenario as to give minimal WER. As noise types noise recorded in a car, white noise and babble, factory 1 and 2 noise taken from the NOISEX database [9] were added.

The results of these audio-visual tests when noise recorded in a car was added can be seen in Fig. 3. The performance of the features based on the energy and on the variance of the coefficients is very similar. Again selecting the features with the highest relative variance seems to work considerably worse. The results obtained with the other noise types are very similar and also indicate comparable performance of the energy and variance features and inferior performance of the features based on the relative variance. They all show that the audio-visual complementarity is well exploited by all feature types and that the audio-visual scores are the better, the better the scores on the video alone is.

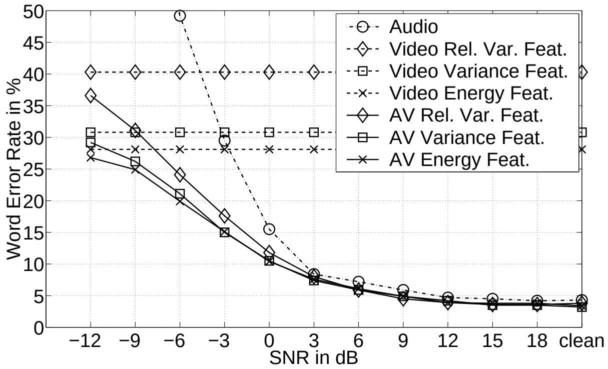

Fig. 3. Comparison of the different DCT feature types. The video only score are also given for orientation

5. DISCUSSION

Throughout this article we investigated different strategies to select DCT coefficients in order to get best recognition scores. During the recording of the database the lightning conditions were kept constant and no normalization of the images was performed. Under these conditions the features based on the energy of the DCT coefficients performed best. The question remains how their performance changes relative to the features based on the variance of the coefficients if the lightning conditions between the training and the test conditions change.

Another interesting question is the difference of performance between the DCT based recognition and one based on geometric lip features as mouth opening or lip shape. One attempt to answer this question is the comparison with a similar database. We previously recorded a single-speaker audio-visual database with a male speaker named John [2] and we therefore want to refer to it as AVNB-John. The current database was recorded with a female speaker named Laurie and hence we want to refer to it as AVNB-Laurie. The lexical content of AVNB-Laurie is a subset of AVNB-John. Feature extraction from the audio stream is almost identical for both databases except for the fact that due to a higher sampling rate at AVNB-Laurie an additive PLP coefficient was used and hence the performance in noise is slightly better. However, as a result of the smaller size of AV-Laurie the performance on clean data is inferior.

The main difference between the two databases is the extraction of the lip features. For AVNB-Laurie the features are extracted via the DCT as detailed before. During the recording of AVNB-John the speakers lips were colored with blue ink and hence the relevant lip parameters could be extracted very easily via a chroma key process. As lips parameters, outer lip width, inner lip width, outer lip height, inner lip height, lip surface and mouth surface surrounded by lips were chosen [10]. By means of these two databases we want to contrast the concepts of pixel based and geometric lip features. In Fig. 4 we plotted the audio-visual recognition scores for the two databases in an identical task. In both cases we used the same phoneme models and dictionary. As can be seen from the plots the two curves representing the audio-visual recognition scores are very similar in shape and coincide at some points. Using the DCT coefficients on AVNB-Laurie we have 28.1% WER compared to 22.1% when using the geometric features on AVNB-John. The feature extraction process from the colored lips comes close to the best achievable precision for geometric features. With natural, uncolored lips the extraction of geometric features is quite delicate. Small variations in the lightning conditions as induced by the articulatory movements can result in severe false localizations of the lips boundaries. Therefore the performance of the geometric lip features in a real life scenario is expected to fall behind those in Fig. 4.b. The video stream in AVNB-Laurie also suffers from the small training set compared to AVNB-John and hence an improvement of the performance of the DCT features can be expected when more training data is available. Due to its simplicity and stability the DCT features can therefore be judged as clearly superior to geometric features. This might also be attributed to the fact that pixel based features are also able to take cues as the visibility of the teeth or tongue and the shape of the muscles around the mouth into account. It has to be said though that in our database the lightning conditions did only change slightly and the position of the lip region could be tracked with a precision of 1 pixel. Both these side conditions favor the DCT. Our results are in accordance with those obtained in [1] where the DCT was compared to geometric features in a continuous density HMM system on video alone.

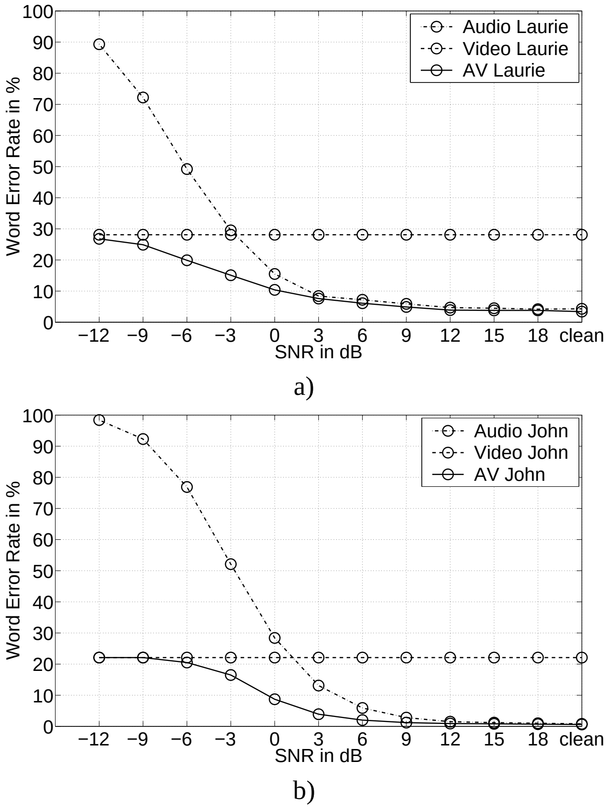

Fig. 4. Confrontation of recognition scores with DCT features (AVNB-Laurie) a) and geometric lip features (AVNB-John) b) on a comparable dataset

6. ACKNOWLEDGMENTS

We want to thank our subjects Laurie Rebut and John Barker for their effort and patience.

7. REFERENCES

[1] G. Potamianos, H. P. Graf, and E. Cosatto, “An image transform approach for hmm based automatic lipreading,” in *Proc. of the Int. Conf. on Image Proc. (ICIP)*, Chicago, 1998.

[2] M. Heckmann, F. Berthommier, C. Savariaux, and K. Kroschel, “Labeling audio-visual speech corpora and training an ann/hmm audio-visual speech recognition system,” in *Proc. of ICSLP 2000*, Beijing, China, 2000.

[3] A. Rogozan and P. Del´eglise, “Adaptive fusion of acoustic and visual sources for automatic spech recognition,” *Speech Communication*, vol. 26, pp. 149–161, 1998.

[4] M. Heckmann, F. Berthommier, and K. Kroschel, “Optimal weighting of posteriors for audio-visual speech recognition,” in *Proc. of ICASSP 2001*, Salt Lake City, Utah, 2001.

[5] University of Mons, Mons, *Step by Step Guide to using the Speech Training and Recognition Unified Tool (STRUT)*, May 1997.

[6] R. A. Cole, T. Noel, L. Lander, and T. Durham, “New telephone speech corpora at cslu,” in *Proc. of EUROSPEECH 95*, 1995, pp. 821–824.

[7] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[8] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, NY, 1993.

[9] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, “The noisex-92 study on the effect of additive noise on automatic speech recognition,” Tech. Rep., Speech Research Unit, Defence Research Agency, Malvern, U.K., 1992.

[10] A. Adjoudani and C. Benoit, “On the integration of auditory and visual parameters in an hmm-based asr,” in *Speachreading by Man and Machine: Models, Systems and Applications*, D.G. Stork and M.E. Hennecke, Eds., Berlin, 1996, NATO ASI Series, pp. 461–472, Springer.