Characterization and Extraction of Mouth Opening Parameters available for Audiovisual Speech Enhancement

Frédéric Berthommier

Institut de la Communication Parlée, INPG
46 Av. Félix Viallet, Grenoble, France
e-mail : bertho@icp.inpg.fr

Abstract

The strong association existing between subbands audio envelope parameters and video parameters extracted using the full DCT (Discrete Cosinus Transform) can be exploited for audiovisual speech enhancement, thanks to a good prediction of amplitude variations by a statistical model. Since the video parameter space is highly multidimensional, the causality of this association must be clarified. At first, a new method of retro-marking is proposed in order to build a transformation function of DCT parameters into explicit ABS mouth opening parameters. Secondly a reduction to single parameter spaces is performed by selection of the best parameters. We show in two noisy conditions that the degradation of the enhancement performance due to the transformation and to the reduction is moderate.

1. Introduction

Recent experimental studies [6, 9] show that audiovisual detection of speech in noise is improved by low level acoustic-visual associations existing between subband envelopes and visible speech cues. The modeling of this low level interaction using linear models proposed by Yehia et al. [11] suggests systematically defining the representations of the two signals leading to a significant linear association. Using the audio visual enhancement paradigm, it has been shown that the audio representation using four subband envelopes (named Sb4) works better than LSP (Line Spectral Pairs) or 16-filterbank representations [2,3]. But, for the video parameterization, this previous study followed the so-called 'pixel-based' approach and used the full DCT format restricted to a high number of parameters (288) only. The aim of the present work is to explicit the video parameters which are available for speech enhancement, which are probably those grounding the perceptive phenomenon of audio-visual coherence.

2. Database and parameter types

2.1 The audio-visual database

The database was recorded for developing an audio-visual speech recognition system [7] based on natural images. This is a repetition of a subset of Numbers95 (OGI) by a single female speaker. The mouth region is well centered and fixed. In the database content, 64*78 RGB images are available after spatial sub-sampling, and at 50ips. The audio is down-sampled from 22KHz to 11KHz. The dataset is composed of about 40000 BMP images taken in the same continuous sequence. The first half is used for training and a part of the second half (about 6 min) for testing.

The sequences are composed of groups (sentences) of English digits and numbers, among 30 different words, separated by silence periods, during which the speaker's mouth is closed. In the present study, two types of noise are used, artificial white noise and natural crowd noise (the same as in [10]). As in [2], a speech interference condition was tested, but the results are less contrasted and then not reported. In order to improve the testing procedure developed in [2,3] the same long test sequence was added with the same noise sequence at different levels, in the [-18,6]dB SNR range, by step of -3 dB, this in the two noise conditions.

2.2 Audio parameters

In order to fit with the classical MFCC representation, the speech signal is decomposed with a Mel-scaled filterbank instead of a Bark-scaled one as in [2,3]. The single audio representation is the spectrally coarse Sb4, using a filterbank composed of four quasi-rectangular filters (Figure 1). To obtain the audio data X, in every 40ms half overlapping hanning window (i.e., at 50 fps), the amplitude levels are calculated in dB RMS.

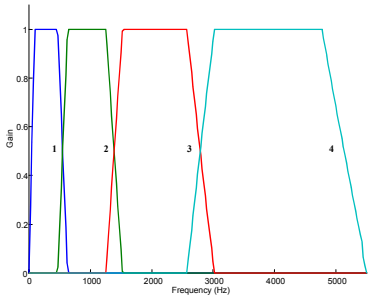

Figure 1: Filterbank design for Sb4. The four quasi-rectangular filters have their high frequency cutoff frequency at: 1) 549 Hz, 2) 1378 Hz, 3) 2783 Hz, 4) 5125 Hz.

2.3 Video parameters

The video parameters are extracted using the full DCT of the initial 64*78 images stored in the database, converted in gray-levels. The audiovisual speech recognition study [7] demonstrated that a small number of DCT parameters, of about thirty, are enough for reaching an optimal video recognition score. These are low frequency domain coefficients, which are distributed more vertically than horizontally. This indicates that the phonetic information is well carried by the small subset of DCT parameters having the largest variability. However, this has the great advantage of the so called pixel-based approach to avoid the use of any face marker. As in the previous studies, the block of first 24*12 (288) DCT parameters have been selected as video parameters, in order to maximize the content of available information, in the DCT condition here considered as a baseline.

Following the geometrical approach of video parameterization, the ABS parameters describe the mouth opening (6 parameters, Figure 2, bottom) and they were classically used at ICP thanks to their excellent properties for carrying explicitly the lip-reading information. Consistently, our first research about the acoustic-visual associations

involved this type of parameterization [1]. Practically, the extraction of ABS parameters from natural images requires a quite complex processing, susceptible to also introduce artifacts. In order to recover ABS parameters from the unmarked images of our video database, a new 'retro-marking' technique is proposed (Figure 2). A transformation function is established relating the DCT and ABS parameter spaces, via a Self Organizing Map (SOM). The SOM algorithm was applied on the training section of the video database, with 288 DCT parameters and 10*10 output vectors. Despite the high number of input parameters, the convergence was remarkably good and easy to reach, showing that the intrinsic dimensionality is very low. Then, the 10*10 components of the bi-dimensional map (SOM DCT) were transformed by inverse DCT and smoothing for visualization. This map is composed, as the input database, of about half of close mouth states corresponding to silent periods. A first axe (horizontal in Figure 2) represents the degree of mouth opening and the second one, vertical, the mouth rounding. The visualization of the 100 components allows drawing by hands on each of them the 8 points needed for defining the 6 ABS parameters and building a second map (SOM ABS, Figure 2). Then the transformation function can be applied over any section of the database because a common SOM label is attributed to each frame using the SOM DCT. The related ABS parameters are selected from the SOM ABS using these labels. Particularly, the train section is marked retroactively (suggesting the name of the method: 'retro-marking').

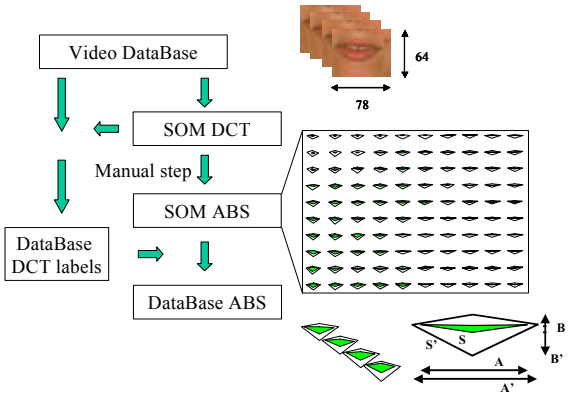

Figure 2: Principle of the retro-marking technique for derivation of the ABS mouth opening parameters from DCT parameters trough a natural image database.

The reduction of the two parametric spaces DCT and ABS, to mono-parametric representations of the video data is also based on the two sets of SOM components because these are representative of the whole database content. In the two cases the reduction is performed by selecting a single parameter in each space with simple heuristics. In the DCT space, a quasi diagonal parameter is preferred because it reflects both vertical and horizontal variations. Then, the linear correlation coefficient between all DCT(i,i) and the 6 ABS parameters is established over the 100 components. The parameters A,A' have the worse correlation pattern. For the ABS space, the inner and outer mouth area S and S' (both calculated by S=A.B/2) are a priori the most informative about the mouth opening state because they include both A and B (resp. A' and B'). Finally, the DCT(5,5) parameter has been selected together with S, and the superimposition of the averaged mouth shape with the related receptive field (Figure 3, COS method) shows its potential ability to well capture the variations of S. Comparatively with the full DCT, the computational cost is greatly reduced since the evaluation of the single output parameter is a multiplication of receptive field values by the gray levels of each image, followed by a summation. An additional simplification of this method for fast video processing substitutes the sign (-1 or 1) of this receptive field, leading to a checkerboard pattern (Figure 3, Check method).

Let remark that the output signal is sensitive to the position of the mouth inside the receptive fields of COS and Check. This has a weak impact in the current framework because the ROI is fixed, but this is a drawback for building a realistic application.

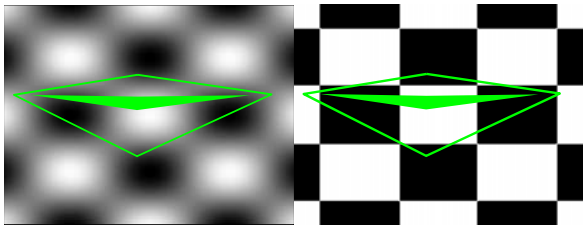

Figure 3: Receptive fields of the COS (left) and Check (right) methods. The average mouth shape is superimposed in the ABS format. The inner mouth area (S parameter) is filled.

Method	DCT	ABS	S	COS	Check
nbp	288	6	1	1	1

Table 1: Number of video parameters (nbp) for the 5 methods.

3. The enhancement process

3.1 Linear estimation stage

According to Yehia et al. [11], a linear statistical model of the audio-visual relationships is built, which allows the prediction of audio data from video data and conversely (see [3]). For each of the 5 methods previously defined, the linear transformation matrix (or vector, for mono-parametric methods) T_{yx} from video data Y to audio data X is estimated from the two time-aligned data sets of the train section (the set of ABS parameters was derived with the retro-marking technique):

$$T_{yx}=(X-\mu_x)(Y-\mu_y)^T((Y-\mu_y)(Y-\mu_y)^T)^{-1}$$

The size of T_{yx} is 4*nbp (nbp, Table 1). The means are calculated over the train section of the database. Then, the estimation of the four audio parameters per frame, at 50 fps, with the video part of the test section (Y), follows a linear rule (in which the audio mean is derived from the train section which is clean):

$$\widetilde{X}=T_{yx}(Y-\mu_y)+\mu_x$$

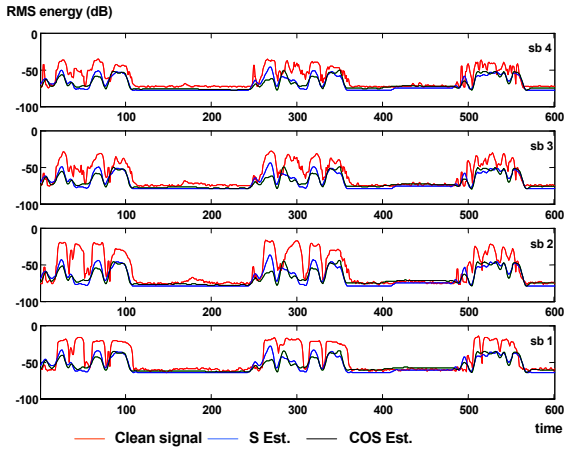

Figure 4: Typical output of the linear estimation process for the S and COS methods.

In Figure 4, a typical output of the linear estimation stage shows that the mono-parametric methods are able to generate estimates which are

temporally correlated with the clean signal envelopes. For the enhancement task, each predicted coefficient is temporally filtered with a 4th order butterworth filter having a cutoff frequency (6.25Hz) above the vocal tract motion range, this is order to reduce the large deviations we observe otherwise in the linear scale. The block diagram of the enhancement process, the same for all methods, is shown Figure 5.

3.2 Wiener filtering of the noisy audio

The Wiener filtering stage consists in decomposing the noisy signal with the four subbands filterbank (Figure 1), and then, the amplitude of the signal arising in each subband time frame is modulated with the related audio estimate. This is by paired multiplication of the linear amplitude values. A reference noted Sb4ref is allowed by the weighting of the noisy signal first decomposed with the quasi-rectangular filterbank, by the linear RMS envelope of the clean signal. In [2,3], the main motivation of having a spectrally coarse Wiener filtering stage was because the spectral information which can be inferred from the video signal is very limited. In the present study, the reduction of the video data to a single parameter (S, COS, Check), together with the use of a linear model, dramatically restricts the prediction in the spectral domain. Particularly, a single video parameter cannot represent the rounding of the mouth as well as ABS parameters (and according to the dual structure of SOM DCT, as DCT parameters). So the spectral features associated with the degree of rounding cannot be predicted. Using the single parameter methods, only the association between the degree of opening of the mouth and the amplitude of the speech envelopes is expected to operate, with the possibility to globally bandpass the speech signal without introducing great spectral distortions.

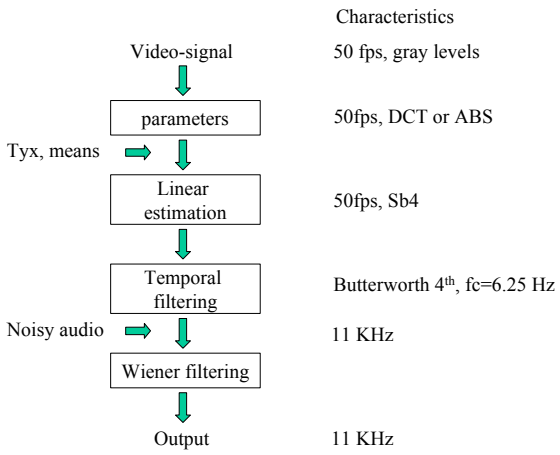

Figure 5: Block diagram and data format of the enhancement process.

4. Evaluation of the enhancement

4.1 Reconstruction accuracy and gain indexes

As in [2,3], we define a spectral distance using the clean speech signal as a reference: the Reconstruction Accuracy (RA) measure. We fix the time-frame duration analysis at 40ms, and the speech silences (defined in clean) are excluded from this statistic. A full-band spectral distance is calculated between the reference R, which is the clean speech, and a signal S. All the spectra are normalized at 1 for removing the effect of global amplitude differences:

$$RA(R,S) = 10 \log \frac{\int_{\Omega} |R(\omega)|^2}{\int_{\Omega} (|R(\omega)| - |S(\omega)|)2}$$

where $\Omega/2\pi = [0,FS/2]$

The RA is an objective index for comparative studies, but the effective gain is the difference between the output RA (S is an output) and the input RA (S' is an input):

Gain= RA(R,S)-RA(R,S')

This spectral gain estimate removes the overall amplitude gain, thanks to the normalization, and it is sensitive to the spectral distortions. Let remark this is a drawback for differentiating the mono-parametric methods, because their audio estimates have by construction similar spectral characteristics. On the other hand, the temporal effect on the overall amplitude is not appreciated. The aim is to preserve the continuity with previous studies.

4.2 Results

For the white noise condition, the RA of the five methods stays between the two references, Sb4ref, built using the clean speech and RA(R,S'), calculated with the noisy speech. The ranking is by average DCT>ABS>S~Check~COS, and it does not vary significantly with the SNR (Figure 6, Table 2). The main observation is that the degradation (relative to the DCT baseline) introduced by the reduction to a single parameter is moderate and close to this produced by the ABS transformation. The three mono-parametric methods have very close performances, but this similarity is probably due to the index itself. The expected equivalence between S and COS methods is better assumed by the generation of correlated estimates (Figure 4), as well as by the direct observation of the high cross-correlation of the video parameters (not shown). Obviously, the receptive fields of COS and Check are similar by construction.

In the crowd noise condition, which is more natural, the ranking of the methods and the pattern of results remains the same, but the gain is not well established because the RA(R,S') is too high in low noise. The quality of the enhancement is also attested by the listening of the output signal in both conditions.

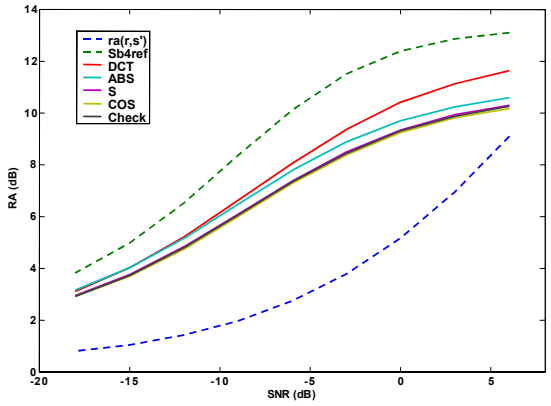

Figure 6: Variation of the RA (in dB) with the input SNR, in the white noise condition.

SNR	Sb4ref	DCT	ABS	S	COS	Check
6	13.11	11.65	10.60	10.31	10.18	10.26
3	12.86	11.15	10.25	9.93	9.82	9.89
0	12.38	10.40	9.71	9.35	9.24	9.31
-3	11.51	9.37	8.90	8.49	8.39	8.46
-6	10.13	8.07	7.79	7.37	7.27	7.33
-9	8.36	6.62	6.48	6.08	6.00	6.05
-12	6.55	5.21	5.17	4.82	4.76	4.80
-15	4.99	4.02	4.03	3.76	3.71	3.74
-18	3.82	3.12	3.15	2.95	2.91	2.93
mean	9.30	7.73	7.34	7.01	6.92	6.97
gain	5.63	4.06	3.67	3.33	3.25	3.30

Table 2: Values of the RA (in dB) in the white noise condition. The last row indicates the average gain.

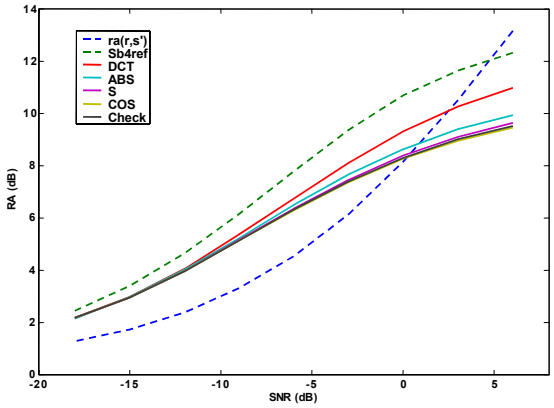

Figure 7: Variation of the RA (in dB) with the input SNR, in the crowd noise condition.

SNR	Sb4ref	DCT	ABS	S	COS	Check
6	12.33	11.00	9.93	9.63	9.46	9.52
3	11.65	10.27	9.39	9.11	8.96	9.01
0	10.68	9.31	8.64	8.40	8.26	8.31
-3	9.37	8.11	7.66	7.47	7.37	7.40
-6	7.80	6.76	6.50	6.36	6.29	6.32
-9	6.16	5.36	5.24	5.15	5.12	5.13
-12	4.65	4.07	4.02	3.98	3.97	3.97
-15	3.39	2.99	2.98	2.96	2.96	2.96
-18	2.44	2.17	2.16	2.17	2.17	2.17
mean	7.61	6.67	6.28	6.14	6.06	6.09

Table 3: Values of the RA (in dB) in the crowd noise condition.

5. Conclusion

We have shown in the context of the audiovisual enhancement paradigm, that the so-called pixel-based and geometric approaches for video parameterization are grounded on the same underlying video features. This modeling study corroborates human experiments [6,9, 10] suggesting that mouth opening (and particularly, area) and subband envelopes of the audio signal are appropriate representational supports of audiovisual coherence and, on the signal processing point of view, of the audiovisual redundancy.

As a straight perspective, the audiovisual enhancement scheme could be improved thanks to a nonlinear model (this was not adapted for a comparison study) as in [1]. Secondly, in the same vein as [5], this model of speech enhancement is a possible front end for ASR applications, working at the signal level (and not at the feature level). This is currently evaluated using the same database and within the framework developed by Heckmann et al. [7]. These results are expected for confirming those obtained with the RA (subject to discussion). Hence, beyond the audio-visual speech enhancement task, the exploitation of the low-level audiovisual coherence offers wide perspectives in the fields of multi-modal signal processing (see [4]), telecommunication and man-machine interaction (e.g., in the MICAL project hosted at ICP). The estimated envelopes are enough temporally correlated with the clean signal (Figure 4) to use this principle in a range of applications related to the extraction of a target speech signal in noise (detection, segmentation) or concurrent speech interference (speaker discrimination, separation). A dual application is audio-visual speaker's mouth localization and modeling of the ventriloquist effect [8]. As a first proposal of implementation in the MICAL platform, which is in line with the present study, the coarse position of the speaker's ROI in a video field could be determined by simple correlation of audio (X) and video (Y) signals; e.g., with the Check structure (Figure 3) for having a fast search. Moreover, using a periodic receptive field structure, this localization estimate might be improved because this is phase sensitive. Let remark that a drawback for the current speech enhancement task (phase sensitivity) could be turned into an advantage for audio visual localization. Further works will specify the coupling between these two applications, enhancement and localization, and the use of amplitude and phase information.

Acknowledgements : This work is a part of a contribution in the MICAL project headed by G. Bailly. I thank L. Rebut, M. Heckmann and C. Savariaux for the elaboration of the audio-visual database. Tables 2 and 3 allow a hyperlink access on samples of the test-set available for listening (if not present, download http://www.icp.inpg.fr/~bertho/ref/bertho-icassp04.pdf).

6. References

[1] Barker, J.P., and Berthommier, F., Estimation of speech acoustics from visual speech features: a comparison of linear and non-linear models, in *Proc. AVSP'99*, Santa Cruz, pp. 112-117, August 1999.

[2] Berthommier, F., Audiovisual Speech Enhancement Based on the Association between Speech Envelope and Video Features, in *Proc. Eurospeech'03*, Geneva, 2003.

[3] Berthommier, F., A phonetically neutral model of the low-level audiovisual interaction, in *Proc. AVSP'03*, St-Jorioz, pp. 89-94, 2003.

[4] Fisher, J.W., and Darrell, T., Informative subspaces for audio-visual processing: High-level function from low-level fusion, in *Proc. ICASSP'02*, pp. 4104-4107, 2002.

[5] Goecke, R., Potamianos, G., and Neti, C., Noisy audio feature enhancement using audio-visual speech data, in *Proc. ICASSP'02*, 2002.

[6] Grant, K.W., and Seitz, P.-F., The use of visible speech cues for improving auditory detection of spoken sentences, *JASA*, 108:1197-1208, 2000.

[7] Heckmann, M., Kroschel, K., Savariaux, C., and Berthommier, F., DCT-Based video features for audio-visual speech recognition, in *Proc. ICSLP'02*, Denver, pp. 1925-1928, 2002.

[8] Hershey, J., and Movellan, J., Using audio visual synchrony to locate sounds, in *Advances in Neural information Processing Systems*, 12, 1999.

[9] Kim, J., and Davis, C., Visible speech cues and auditory detection of spoken sentences: an effect of degree of correlation between acoustic and visual properties, in *Proc. AVSP'01*, Aalborg, pp. 127-131, 2001.

[10] Schwartz, J.-L., Berthommier, F., and Savariaux, C., Auditory syllabic identification enhanced by non-informative visible speech, in *Proc. AVSP'03*, St-Jorioz, pp. 19-24, 2003.

[11] Yehia, H., Rubin, P., and Vatikiotis-Bateson, E., Quantitative association of vocal tract and facial behavior, *Speech Communication*, 26(1):23-43, 1998.