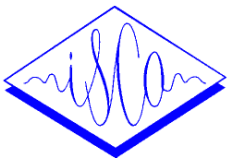

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

A Phonetically Neutral Model of the Low-level Audiovisual Interaction

Frédéric Berthommier

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

Abstract

The improvement of detectability by visible speech cues found by Grant and Seitz (*JASA*, 108:1197-1208, 2000) has been related to the degree of correlation between acoustic envelopes and visible movements. This suggests that the audio and visual signals could interact early during the audio-visual perceptual process on the basis of audio envelope cues. On the other hand, acoustic-visual correlations were previously reported by Yehia et al. (*Speech Communication*, 26(1):23-43, 1998). Taking into account these two main facts, the problem of extraction of the redundant audio-visual components is revisited: The video parametrization of natural images and three types of audio parameters are tested together, leading to new and realistic applications in video synthesis and audiovisual speech enhancement. Consistently with Grant and Seitz' prediction, the 4-subbands envelope energy features are found to be optimal for encoding the redundant components available for the enhancement task. The computational model of audio-visual interaction which is proposed is based on the product, in the audio pathway, between the time-aligned audio envelopes and video-predicted envelopes. This interaction scheme is shown to be phonetically neutral, so that it will not bias the phonetic identification. Then, the low-level stage which is described is compatible with a late integration process, and this is a potential front-end for speech recognition applications.

1. Introduction

The perception of speech is greatly improved in the presence of visual information, the mouth movements and the talking face, and a gain of intelligibility of about 10-15dB is classically reported. In a seminal paper, Summerfield [18] analyzed the origin of this gain, and the potential roles of the visual cues. At first, these can provide complementary information about the place of articulation, which is the most degraded in the audio signal, and, at the same time, the easiest to lip-read. A great part of the literature focused on this main property, and the other possible factors attracted little interest. Using a detection paradigm of speech in loud noise, Grant and Seitz [11] assessed evidence for another mechanism evoked by Summerfield [18], which is based on the temporal coherence between lip movements and the speech envelope cues. In the audiovisual (AV) condition, a release of masking of about 1.6dB was found, relative to the audio only (AO) condition. These results were confirmed by Kim and Davis [14] with a larger dataset. This facilitation of the detection of the speech segments near the threshold was attributed to the linear correlation existing between the mouth aperture and the energy envelope of speech, overall and decomposed in subbands. A better correlation was found in the 2nd and 3rd formant regions, consistent with the speechreaders' ability to extract the place of articulation. Although the level of this facilitation is presumably early, the role of the temporal coherence was not considered as apart from this of the phonetic complementarity.

The functional specificity of this enhancement process was revealed by an articulatory-feature detection paradigm proposed by Schwartz et al. [16] (renewed after the failure of Barker et al. [1] for retrieving an effect with the {/d/,/g/} contrast). This shows that the near threshold transmission of the voicing feature is facilitated in the AV relative to the AO condition. Hence, the intervention of the phonetic audio-visual complementarity is discarded because the voicing cue is completely absent in the visual information. Remarkably, in this experiment, an identification gain was directly measured, despite the near threshold characteristic, and the origin of this gain was identified. The level of the interaction appears clearly as pre-phonetic: In a detection-identification pathway, an audibility gain, corresponding to a detection improvement at the feature level, leads to an intelligibility gain due to a better phonetic identification.

These experiments ([11],[14],[16]) well establish that an audiovisual interaction operates early, and before the phonetic integration. Since these are basic detection tasks, very little is apparent about the detail of the process, necessary to know before building up a computational model. To guide us, in their recent review developing many parallels between functional and neurophysiological data, Bernstein et al. [3] distinguish AV interactions that result from information processing (i.e., integration) and those that just modulate the activity levels. Strictly speaking, if we consider that the detection is inherent in the process itself, the AV interaction participates to the information processing. So, there is an alternative position in which the detection is the task (i.e., the observable) instead of being a function. The goal of this paper is to propose a type of interaction which is modulatory. Then, the main property to establish is the *phonetic neutrality*. To fulfill this condition, the interaction does not bias the phonetic identification.

As shown by the old Erber's experiments [8], the speech envelope cues carried by the overall RMS energy are weakly intelligible in isolation, but complementary to lip-reading cues. When the spectral reduction is not complete, thanks to subband decomposition, the speech intelligibility is remarkably increased with just four subband envelopes modulating white noise [17]. In this case, the audiovisual speech complementarity operates effectively, leading to an almost perfect intelligibility, because the place of articulation is not well transmitted by the AO [4], whereas the voicing and the manner are well represented. This is consistent with the blurring of the formant structures (i.e., peaks and trajectories) in the spectrally reduced speech (SRS). However, some perceptible place-of-articulation cues are present, which are easy to identify in the acoustic signal, as the burst of the plosives /g/ and /k/. The choice of this representation for modeling a modulatory AV interaction was practically initiated by convergent previous works ([2],[4]), as well as by related 'CASA style' modeling. However, in the current framework, this is motivated **(1)** by the finding of correlations between the mouth aperture and the energetic envelope in subbands [11], **(2)** by the coarse spectral and amplitude modulation characteristics of the filtered envelopes which are compatible with a low-level processing as well as by **(3)** its relative phonetic neutrality, due to the spectral reduction.

The last question to address for completing the basis of a computational model concerns the type of AV transformation which is implicated. This was pointed out by Grant and Seitz [11]: "*Exactly how much information about the temporal and spectral envelope can*

be gleaned via speechreading is not clear, although a recent study by Yehia et al. [19] suggests that 70%-80% of the variance in the rms amplitude can be recovered by nonlinear transformations of facial motion". The Yehia et al. study [19] and further confirmations ([2],[13]) reported a significant association between acoustic features (Line Spectral Pairs + overall RMS energy) and the position of facial markers. This association can be captured with linear transformations, after a frame by frame training with audiovisual data without any labeling. Then, this is possible to predict a part of the markers' position information from the audio signal and conversely.

2. Parameter types

In preliminary works ([5],[6]), the observation of the same acoustic-visual linear association was extended for other parameter types and the use of facial markers was avoided. Then, the feasibility of two main applications, video synthesis and speech enhancement, was tested. We develop in this paragraph the important aspects, and point the reader to these papers for completing the technical details. For the video synthesis from speech sounds only [5], this association was established between the video DCT (Discrete Cosine Transform) features and the audio DCT parameters, equivalent to the MFCC (Mel Frequency Cepstral Coefficients). The video parameters were extracted from a database of natural images of the region of interest (ROI), at first used for a recognition application [12]. The video recorded at 50ips was reduced and the centering on the mouth region has been stabilized.

In order to extend the Yehia et al. [19] approach, we have shown that a database large enough allows to train the two linear associators between image and sound, Txy and Tyx, this from natural images parametrized with a small number (24*12=288) of DCT coefficients. For achieving a synthesis application the advantage of the DCT is to be reversible. Hence the synthetic images generated with the audio-to-video associator Txy are small and blurred, but their appearance is natural. There are a few artifacts in the reconstruction of the lips movements, and the main observation is that the dynamic is globally hypo-articulated. Moreover, the independent processing of the RGB colors allows the reconstruction of naturally colored images.

The second potential application is audio-visual speech enhancement, using the video-to-audio associator Tyx. A method for audio speech enhancement from geometrical lips parameters was proposed by Girin et al. [9]. This is based on spectral estimation and Wiener filtering: For each time frame, a filter is estimated from the video parameters and then applied on the noisy audio signal. However, the approximated filter is expected to carry rather precise spectral information about the target speech. In the context of speech recognition, Goecke et al. [10], proposed to enhance the audio features instead of the audio signal.

For the audiovisual speech enhancement in loud noise and speech interference [6], three types of audio parameters were tested, including the initial LSP+RMS, the DCT and the four subbands RMS energy (Sb4). Let remark that, in [5], the use of the DCT on the audio side was motivated by the symmetry of the design, and in [6], by the possibility to reduce the dimension to four coefficients (DCT4). An advantage was found for the envelope features (Sb4), as well as for this reduced form. Because only linear transformations are applied, the full DCT method based on 16 coefficients is formally equivalent to the direct use of the 16 RMS envelopes at the output of the filterbank, and this could be named Sb16. The video-to-audio prediction of these coefficients, frame by frame, corresponds to a spectral evaluation which is not constrained by an a priori about the format of the AV redundant components. Their application as a Wiener filtering for speech enhancement is neutral. Hence, the counterpart is the dispersion of the information among the 16 dimensions.

The choice of the audio parameter space is critical for the observation of acoustic-visual correlations, and their interpretation. Hence, these are the consequence of the physical dependence between mouth/face configurations and vocal tract configurations (i.e., between visible and non visible speech movements), but the presence of acoustic correlates can be observed with different point of views. The optimal parameters for capturing this association are a priori those encoding the formant configurations and, better, those estimating the formant trajectories. This motivated the choice of the Line Spectral Pairs as acoustic parameters, with an additive RMS parameter for representing the overall amplitude. In a speech enhancement application, this is also required for evaluating the gain of the related linear filter. The LSP method is the most constrained, and, for speech enhancement, the counterpart is the introduction of distortions, because the frame by frame estimate with a global linear associator is not precise.

The encoding of the speech in 4-subbands envelope levels (Sb4) per frame corresponds to the spectral reduction of the speech [17]. With quasi-rectangular filters (Figure 1), the spectral structure of the speech is flattened and the residual speech cues are mainly represented by slow amplitude modulations along the temporal axis. This strategy is orthogonal to that of the LSP method, and this is consistent with the a priori that the temporal coherence is an important factor of AV coherence. The AV redundant components are represented in the temporal domain. For speech enhancement, the application of flat filters is neutral. Hence, the counterpart is the impossibility to predict fine spectral structures and to add information to the audio thanks to the Wiener filtering.

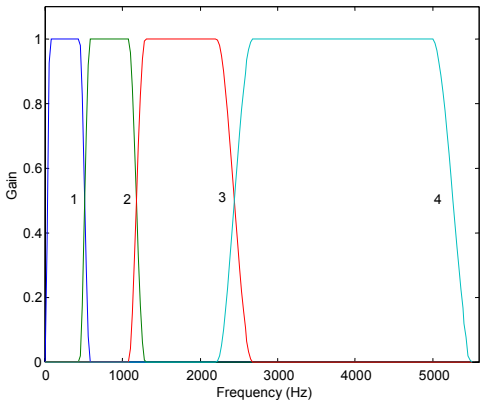

Figure 1: Filterbank design for Sb4. The four Barkscaled and quasi-rectangular filters have their high frequency cutoff frequency at: 1) 515 Hz, 2) 1175 Hz, 3) 2440 Hz, 4) 5250 Hz. The audio signal is sampled at 11025Hz.

3. Video synthesis

3.1 Method

The linear transformation matrix T_{xy} from audio data X to video data Y is estimated from the AV synchronous data of the training section of the database (about 20000 frames):

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

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

In a second stage, the prediction of the 288 DCT coefficients per frame is performed at 50fps and each coefficient is temporally filtered with a butterworth having a very low cutoff frequency (3.25Hz) in the upper bound of the vocal tract motion range. For the synthesis of RGB sequences, this process is repeated three times, for each color component independently (for the other technical details see [5]). In the present study, we incorporate the three audio formats (LSP, DCT, Sb4) described before.

Figure 2: Three images generated from speech sound, with the 3 parameter types (LSP,DCT,Sb4), in comparison with the synchronous original (ORI). For homogeneity, this is processed by compression to 288 DCT coefficients and interpolation.

A qualitative comparison of the video sequences generated by the 3 methods shows that LSP and DCT perform similarly, whereas the generation from subband levels (Sb4) carries less lip reading cues. The mouth opening is always represented (Figure 2, compare to ORI). Hence, a lead of 2-4 images exists between the original video (ORI) and the synthetic sequences.

3.2 Intelligibility gain in loud noise

This perceptual experiment is an adaptation of an automatic speech recognition paradigm for quantifying the human audio-visual gain of intelligibility in noise allowed by the generated video sequences. This was designed for the evaluation of the video synthesis application [5]. Depending on the content of the training database, the task is to identify English numbers in crowd background noise. The sequences are composed of concatenated sentences of English digits and numbers, with a vocabulary of 30 different words, each sequence including 31 target digits by average. A set of 9 noisy AO sequences in the [-18,6]dB SNR range is presented at first in low noise and the SNR is progressively degraded by step of -3 dB. The AV part is composed of 8 sequences, each generated from the DCT parameters of the clean audio speech of the test section of the database. Then, the audio signal is added with crowd noise in the [-21,-3]dB range with the same step of SNR, and the last sequence is video alone. Audio and video channels are mounted with *Adobe premiere* with an audio lag of 66ms.

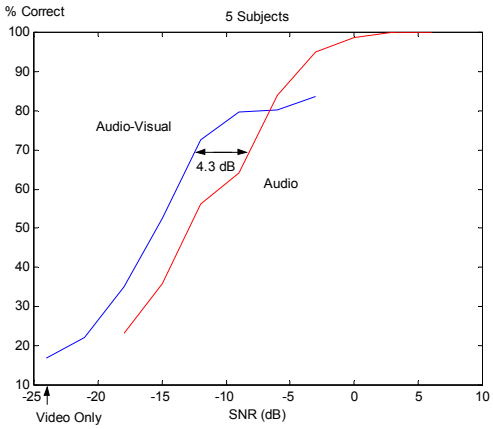

Figure 3: Measure of the intelligibility gain for five subjects.

The responses were written by the listeners on a paper form indicating the structure of each sentence. The score is expressed in percent correct digit identification without realignment. The numbers are decomposed in digits. The results of five subjects have been averaged. In Figure 3, we observe a steep degradation of the audio intelligibility below -3dB SNR. The intelligibility gain is quantified by the maximum deviation between the two curves, AO and AV, which is 4.3 dB. The deviation is rather constant and close to 2.5dB for the lower SNR levels, towards the speech detection threshold, allowing a score improvement of about 15%.

This gain is a fraction of the AV improvement expected using the original images, which is allowed by only the cues which are present after the audio-to-video transformation. Since the audio-parametrization is based on the DCT, some residual phonetic cues are encoded in the video signal. A smaller gain is expected with Sb4, in the same range as the detection gain found by Grant and Seitz [11].

4. Audio-visual speech enhancement and synthesis

4.1 Method

Symmetrically to the video synthesis, the linear transformation matrix T_{yx} from video data Y to audio data X is estimated from the synchronous frames, audio and video, of the training section of the database:

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

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

The audio frame duration is 40ms, half-overlapping, and the three types of predicted parameters for each frame are Sb4 (nbp=4), LSP (nbp=24+1), DCT (nbp=16). In all cases, these coefficients are filtered temporally with a 4th order butterworth filter having a cutoff frequency about twice this adopted for the video synthesis (6.25Hz). This is in order to reduce the large deviations we observe otherwise.

These estimated audio parameters are used in two conditions:

- Condition **(1)**: For speech enhancement, a Wiener filtering is applied on the input (noisy audio) data. The linear filter is directly derived from the estimated values for the LSP method, and using the filterbank structure for the DCT and Sb4 methods; e.g. for Sb4, the input signal is decomposed in 4 subbands by the filterbank (Figure 1), and frame by frame, the amplitude in each subband is multiplied by the value of the related parameter. The output signal is recomposed by simple summation.

- Condition **(2)**: For sound synthesis from video, this is the same method, but the input signal is white gaussian noise (this is the noise only condition, NO). For Sb4, this leads to the prediction of SRS from video.

Then, the Sb4ref reference is generated from the clean speech, thanks to the same 4-subbands decomposition (Figure 1). The envelopes of Sb4ref are the RMS levels in each 40 ms frame. For each level of the condition (1), these envelope levels are multiplied with the envelopes of the input signal, and for condition (2), these modulate white noise, and this is strictly equivalent to SRS synthesis. Sb4ref is also the target of the Sb4 prediction.

The first condition was tested in [6], but the input speech was corrupted with crowd noise and speech interference. In order to set continuity between conditions (1) and (2) the same simulation is repeated, but the input speech is corrupted with white noise. In

condition (1), the noise progressively replaces the speech signal, and the limit is the condition (2).

4.2 Simulations

The 9 sequences of the test database are corrupted with white noise in the [-18,6]dB SNR range for the speech enhancement test. For the condition (2), only the video channel of the same 9 sequences is used. The index of Reconstruction Accuracy (RA) is applied for evaluating the quality of the enhancement (see [6] for more details):

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

where R is the reference (clean speech), and S the signal to test (silence excluded). The effective enhancement gain (which has no sense in condition (2)) is the difference between the output RA (S is an output) and the input RA (S' is an input):

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

SNR	ra(r,s')	Sb4ref	Sb4	LSP	DCT
6	8.14	12.51	10.98	6.31	8.30
3	6.76	12.74	11.46	6.67	8.43
0	5.00	12.30	10.25	6.82	7.91
-3	3.76	11.25	8.67	6.22	6.79
-6	2.78	9.81	7.57	5.88	6.11
-9	2.11	8.73	6.70	5.97	5.67
-12	1.42	6.88	5.34	5.40	4.78
-15	1.16	5.41	3.88	4.20	3.10
-18	0.86	3.98	3.16	3.46	2.83
mean	3.56	9.29	7.56	5.66	5.99
gain	-	5.74	4.00	2.10	2.44
NO(seq7)	-	1.77	1.51	1.98	1.43
NO(mean)	-	1.71	1.43	1.79	1.32

Table 1: Values of the RA (in dB) in the white noise interference condition. In the second column, ra(r,s') is the RA of the input signal. The last two rows indicate the RA in the condition (2) (small fluctuations of all these measures are observed, due to the use of white noise).

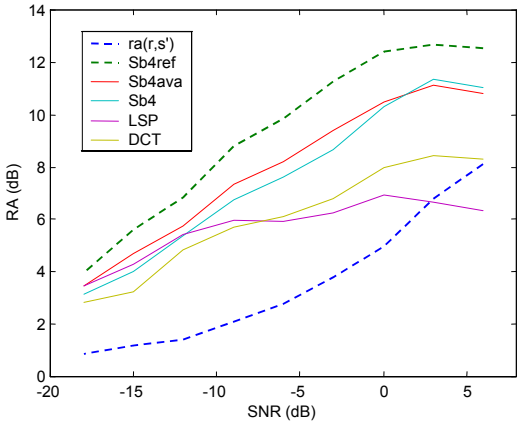

Figure 4: Graphical results of the condition (1), partly shared with Table 1.

In condition **(1)**, the optimality of the Sb4 method above -12dB is attested, and the use of LSP for speech enhancement is clearly worse. This confirms [6], because the profiles of the three curves (LSP, DCT, Sb4) are similar, and, moreover, the two references Sb4ref and ra(r,s') are better set as upper and lower boundaries (Table 1, Figure 4). Conversely, in condition **(2)**, the optimal speech synthesis is, obtained with the LSP method (Table 1, Figure 5).

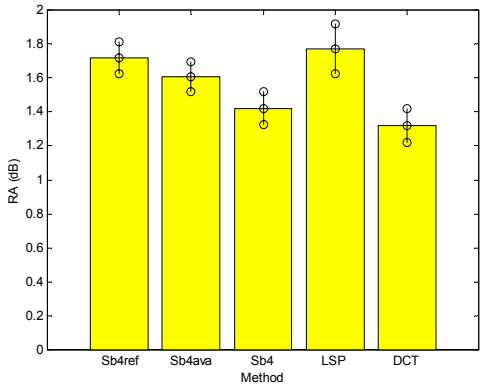

Figure 5: Results of the condition (2). The error bars indicate the std between the 9 sequences.

These two observations lead to the conclusion that the predicted LSP parameters carry more speech information. But this is not appropriate for speech enhancement because it distorts the audio speech features which are not masked by the noise. On the contrary, the flatness of the Sb4 Wiener filtering confers spectral neutrality to this method, and then better gains for moderate noise levels.

5. Assessing the phonetic neutrality of the interaction

5.1 Motivation

The prediction of the clean speech envelopes by Sb4 is quite good (see [6]), and the predicted SRS is partly intelligible. For assessing the complete neutrality of the modulation, it is necessary to show that the predicted envelope does not bias the audio signal in the temporal domain (as the LSP method does in the spectral domain). This is not taken into account by the RA index, which is essentially a spectral distance. As mentioned in the introduction, the SRS carries residual place-of-articulation cues. We will show that these cues are not transmitted via the predicted envelope from the visual pathway to the audio pathway. The hypothesis is that the proposed interaction cannot produce a bias, as that of the McGurk effect, which is well observed with the SRS [4].

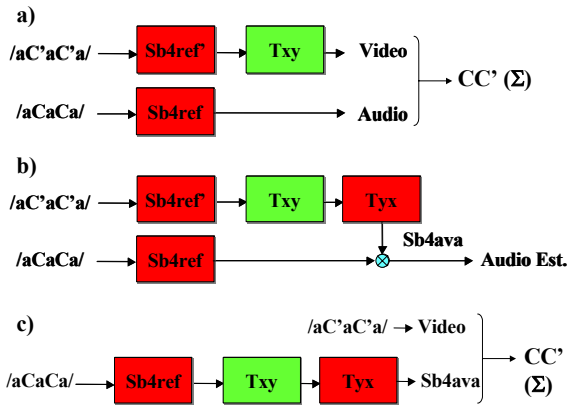

Figure 6: Three tests for assessing the phonetic neutrality of the interaction in the temporal domain (See text).

Using /aCaCa/ utterances extracted from the corpus of [4], three conditions are elaborated in order to test this hypothesis (Figure 6): **(a)** Coherent and non coherent AV pairs are composed, with audio SRS, and video synthesized with the Sb4 Txy associator (which uses

Sb4ref' parameters as input). These pairs are submitted to subjects in a perceptual experiment. **(b)** The AV interaction produced by these AV pairs is predicted, and an audio estimate which is the product of the interaction is synthesized. **(c)** Non coherent AV pairs are built with the predicted envelopes and the original video sequences.

5.2 Testing of the phonetic neutrality

For the test **(a)**, four utterances of each AV pair /bb/,/bg/,/db/,/dg/ are built, and a sequence of 16 stimuli /aCaCa/ was submitted to five subjects with 3 forced choices per consonant. The responses were written, with the possibility to differentiate the first and the second consonant of each stimulus (40 judgments/block). The confusion matrix Figure 7 shows that the responses are not biased by the synthesized video: there is no significant difference between /bb/, /bg/ and between /db/, /dg/. This failure of observing a McGurk effect is not surprising after comparison of the two video sequences which are synthesized from the audio /b/ and /g/: the sequences are very similar, and the mouth movement is intermediate. As expected, the frame by frame linear associator cannot capture the temporal events present in the envelope signal, and convert them in appropriate mouth movements. Hence, this conclusion is attenuated by the use of an associator which is trained on a very different corpus.

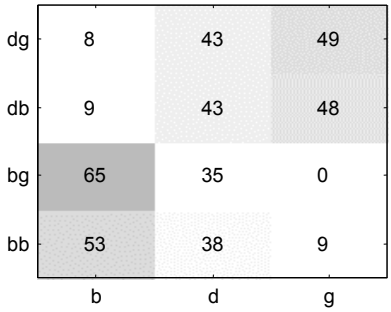

Figure 7: Confusion matrix of the test (a), expressed in percentages.

The test **(b)** is based on the prediction of the interaction which occurred in the test **(a)**, and of the audio signal which is supposed to be perceived by the subjects after this interaction. The video-predicted envelope, named Sb4ava is generated by application of the two associators on the envelope parameters:

$$\vec{\vec{X}} = T_{yx} T_{xy} (X - \mu_x) + \mu_x$$

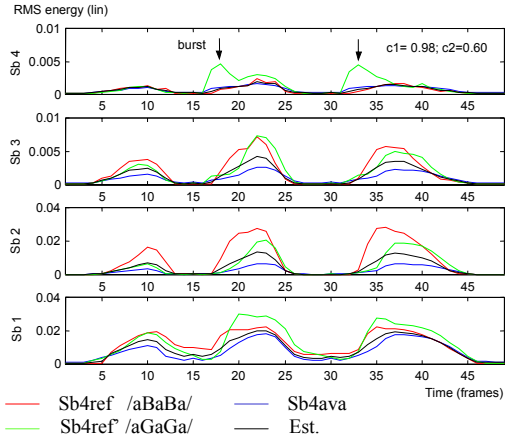

Figure 8: Test (b) simulation of the /bg/ pair. The burst of the /g/ is pointed in the subband 4 (Sb 4). The square root of the estimated envelope (Est.) is plotted and compared (for Sb 4 only) by linear correlation, to /b/ (c1=0.98) and /g/ (c2=0.60) envelopes. This is closer to /b/.

Then, Sb4ava modulates the (SRS) input signal in order to produce an estimated audio signal. At first, this method has been applied for speech enhancement and synthesis, and the results plotted in Figures 4 and 5. These are equivalent to those of Sb4, which are based on direct predictions. In the figure 8, the four envelopes involved in the /bg/ pairing are displayed. The burst of /g/, carrying the place-of-articulation contrast, is visible in the fourth subband, whereas there is no burst for /b/. Because the predicted envelope is flat and does not carry the burst, the product between the envelope of /b/ and this predicted envelope is similar to /b/, and then, the estimated audio is closer to /b/ (Figure 8). This can be confused with /d/, but rarely with /g/ (as in the experiments with audio SRS, see [4]).

The test **(c)** consists in synthesizing speech with the video-predicted envelope and then pairing with the original video. An informal experiment shows that the response is determined by the video only. This leads to a normal McGurk effect /bd/ or /bg/>/d/, as well as to an inverse McGurk /db/ or /gb/>/b/. The phonetic neutrality of these audio signals is explained by the structure of the transformation matrix Tyx*Txy which serves for their generation (Figure 9). Through this transformation, we find a strong interaction between the 2nd subband and the others. At the same time, for SRS, the 2nd subband is the one carrying the less phonetic cues (the 1st carries the voicing cue, and the 3rd,4th carry mode and place of articulation). Then, the temporal phonetic cues of the SRS are filtered out by this transformation, and the resultant interacting envelopes are phonetically neutral.

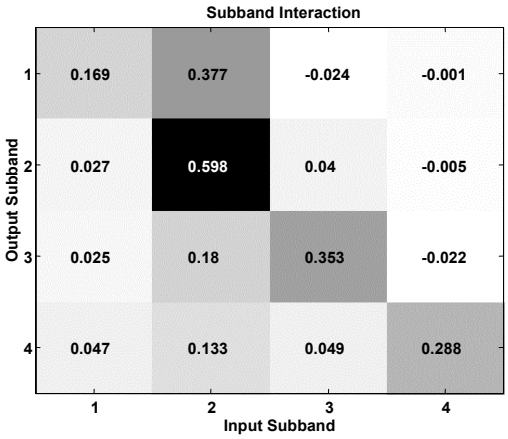

Figure 9: Subband interaction through the Tyx*Txy transformation.

6. Conclusion

Following this model, one role of the low-level interaction is to reinforce the amplitude modulation of the speech segments, this without distortion of the phonetic cues, spectral or temporal. This could explain a speech detection improvement at the threshold level, and at the supra-threshold level, an intelligibility gain due to the visual cues. For applications, the property of phonetic neutrality allows us to use the model as an enhancement front-end for an audio speech recognition process, or an audiovisual recognition system.

In the perspective of modelling the audiovisual integration, other potential roles can be attributed to AV interactions preceding the phonetic identification. The late integration model proposed by Massaro [15] has been extended with a low level stage able to extract information about the synchronisation of the two channels, which is addressed to the integrative stage as a supplementary cue. In the current framework, this is possible to perform a synchrony measure (e.g., of the delay) between the audio envelopes and the internal envelope generated from the visual signal, via the transformation Tyx. Moreover, the low-level interactions are possible basis for binding the audio and the visual streams. In the theoretical framework developed

by Bregman [7], this type of interaction can be qualified as primitive, and this is clearly differentiated with binding/grouping processes appealing memorised information (called 'schema-based'), corresponding here with the upper level of phonetic coherence. Hence, this correspondence needs caution because these concepts have been adapted for the auditory scene analysis. Here, the 'primitive levels' are probably situated in speech specific modules. But this is helpful to understand what is the function of 'low-level' audio visual processing, and, in this way, we have built a simple demonstration showing an effect of AV primitive grouping. One example of auditory streaming involving speech components is the perception of the clicks of the !Xoo language, which are heard separately. The coherence of the speech stream can be restored using a synthetic video generated with Txy, because the click is associated with a small movement of the mouth. Then, the two elementary events are synchronous, and these are grouped together in the same perceptive AV stream. Moreover, the continuity of the visual cues allows the grouping with the other phonetic components. So the binding between audio and visual information could also be primitive and pre-phonetic.

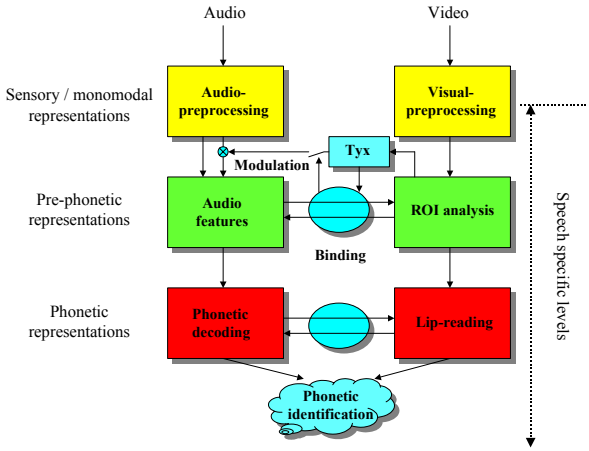

Figure 10: Incorporation of the low-level interaction in a modality specific architecture [3], in which there is both an audio and a video phonetic processing stream.

In figure 10, the incorporation of the modulatory interaction is designed to be compatible with a modality specific (M-S) architecture and it is inspired by the neurophysiological data and the framework detailed in [3]. The four subbands decomposition is placed in a pre-processing module which is partially monomodal. We also remark that the interaction could be controlled and established by the binding operator itself, since it depends on the presence of speech sounds related to mouth movements at a pre-phonetic level. Then, the two relationships with the low-level binding module are (1) a capacity to perform a synchrony measure and (2) the on-off control of the modulation. The interaction itself is produced by a descending connection. Anatomically, this interaction could be placed between the STG and the STS. The phonetic integration occurs later in this pathway, and this architecture allows room for a second/repetitive node placed between the second and the third stage, for the implementation of non neutral interactions, as these described by Girin et al. [9].

Acknowledgements : This work is a part of the CTI-STIC project "Etude psychoacoustique et modélisation computationnelle des mécanismes de décodage acoustico-phonétiques à partir de la parole dégradée spectralement et temporellement". I thank L. Rebut, M. Heckmann and C. Savariaux for the elaboration of the audio-visual database, R. Colleti for his assistance during the summer 2002.

7. References

[1] Barker, J.P., Berthommier, F., and Schwartz, J.-L., Is primitive AV coherence an aid to segment the scene ?, in *Proc. AVSP'98*, Terrigal, 1998.

[2] 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.

[3] Bernstein, L.E., Auer, E.T., and Moore, J.K., Audiovisual speech binding: Convergence or association ?, in *Handbook of Multisensory processes*, Calvert, G., et al. (Eds), Cambridge, MIT Press (in press).

[4] Berthommier, F., Audio-visual recognition of spectrally reduced speech, in *Proc. AVSP'01*, Aalborg, pp. 183-188, 2001.

[5] Berthommier, F., Direct synthesis of video from speech sounds for new telecommunication applications, in *Proc. SOC'03*, Grenoble, May 2003.

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

[7] Bregman, A.S., Auditory scene analysis, Cambridge, Mass, MIT Press, 1990.

[8] Erber, N.P., Speech-envelope cues as an acoustical aid to lipreading for profoundly deaf children, *JASA*, 51:1224-1227, 1972.

[9] Girin, L., Schwartz, J.L., and Feng., G., Audio-visual enhancement of speech in noise, *JASA*, 109(6):3007-3020, 2001.

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

[11] 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.

[12] 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.

[13] Jiang, J., Alwan, A., Keating, P. A., Auer, E.T., and Bernstein, L.E., On the relationship between face movements, tongue movements, and speech acoustics, *EURASIP JASP*, 11, 1174-1188, 2002.

[14] 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.

[15] Massaro, D.W., Perceiving talking faces: From speech perception to a behavioral principle, MIT press, Cambridge, MA, 1998.

[16] Schwartz, J.-L., Berthommier, F., and Savariaux, C., Audio-visual scene analysis: evidence for a 'very-early' integration process in audio-visual speech perception, in *Proc. ICSLP'02*, Denver, pp. 1937-1940, 2002.

[17] Shannon, R. V., Zeng, F-G., Kamath, V., Wygonsky, J., and Ekelid, M., Speech recognition with primarily temporal cues, *Science*, 270, 303-304, 1995.

[18] Summerfield, Q., Some preliminaries to a comprehensive account of audio-visual speech perception, in *Hearing by Eye: The psychology of lip-reading*, Dodd, B. and Campbell, R. (Eds.), Lawrence Erlbaum, Hillsdale, 1987.

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