Audio-visual recognition of spectrally reduced speech

Frédéric Berthommier

ICP/INPG, 46 Av. Félix Viallet, 38031 Grenoble Cedex
fax : +33476574710 and e-mail: bertho@icp.inpg.fr

Abstract

Perceptual experiments on audio-visual consonant recognition based on the spectral reduction of the speech (SRS) have been carried out with coherent and incoherent (McGurk) audio-visual pairs. The main interest of SRS in four sub-bands is to have a partial suppression of the information transmitted for the place of articulation. The integration of manner, restricted to the fricative/occlusive contrast, is also of concern, and a new 'cross-manner' combination is tested. As expected, we have a good audio-visual complementarity for SRS and a high amount of McGurk responses, but new interesting effects are observed. For the interpretation of human confusion about place of articulation, the Bayesian model proposed by Massaro and Stork [8] is compared to a new place identification model which is based on averaging as well as on the separate identification of articulatory features. This decomposition is a promising way for the development of multi-stream speech recognition models.

1. INTRODUCTION

In an initial experiment, Erber [3] used the technique proposed by Horii et al. [7] to show an improvement of the audio-visual intelligibility when the audio speech signal is reduced to its temporal envelope, relative to the video only condition. The Horii's technique consists of extracting the envelope of the fullband wave and then modulating white noise. Despite the complete absence of spectral information in this signal, and its very low intelligibility, this demonstrates that the temporal envelope carries cues complementary to video. More recently, with the same paradigm, van Tassel et al. [13] varied the temporal resolution of the temporal envelope. Moreover, they better quantified the transmission of phonetic information as defined by Miller and Nicely [10] in terms of articulatory feature transmission (voicing, manner and place). Later, the degree of spectral resolution of envelope speech was controlled by Shannon et al. [12] by dividing the spectrum in a variable number of subbands, instead of using the fullband envelope as in previous studies. When they varied the number of subbands from one to four, the transmission of phonetic information was restored at four subbands, except for the place of articulation, which was only partially recovered at this resolution.

The purpose of this paper is to use spectrally reduced speech (SRS) as described by [12] to carry out experiments on audiovisual speech perception. The task is limited to consonant speech recognition. Using such a signal, the place of articulation is partially transmitted by the audio, and this could complement the observations made with noisy speech by Sekiyama and Tohkura [11] who found an enhancement of the classical McGurk effect [9]. Interestingly, for audiovisual SRS, the fricative/occlusive contrast, which is a manner characteristic, plays an important role, and we experiment new audio-visual 'cross-manner' combinations, as well as a noisy condition, which operates on manner transmission.

Since the challenge is to find robust algorithms for multistream speech recognition[1], we will discuss the problem of the modelling of audio-visual fusion in light of these experiments. Then, we will test the prediction of human AV responses from audio only (AO) and video only (VO) responses, including the prediction of audio-visual pairs, which are not compatible. In this way, two main proposals have been developed by authors who consider that audiovisual identification is the result of a fusion process which takes into account both audio and video. The first one is proposed by Braida et al. [2] and it relies on a response center (prototype) expressed in a multidimensional representation, where the articulatory features are not distinguished. The second method is proposed by Massaro and Stork [8] who apply the multiplicative Bayes rule at the class identification level, to predict human responses as well as to implement automatic recognition.

2. DATABASE AND EXPERIMENT SETUP

2.1 Database and signal processing

A small audio-visual database of 48 aCaCa utterances (12 French consonants and four repetitions of each) pronounced by a single male speaker were recorded with a CANON MV20i digital camera. The data were segmented in one-second duration video files at 25 frames/s under *Adobe premiere*. The synchronised audio tracks sampled at 11025 Hz were processed separately and then dubbed. To perform the spectral reduction of the speech, a technique similar to this described by [12] was applied. The four filters are Bark-scaled and quasi-rectangular (Fig. 1). The subband waves arising from these filters are demodulated by half-wave rectification and 1st order Butterworth filtering, with a cutoff frequency at 10 Hz or 500 Hz (this defines two conditions). A white noise is modulated by the four resultant envelopes and the stimulus is recomposed by summation. In the noisy condition, an AM noise is added at 3dB global SNR. The main characteristic of SRS (when the number of subbands is low) is the loss of harmonicity as well as of the formantic structures (formant shape and formant trajectories).

2.2 Choice of the set of stimuli

In the experiments presented in this paper, we take 12 French consonants {pbfvtdszkgʃʒ} in order to have a classification task using systematically the 3 dimensions of voicing, manner and place (Table 1): Voicing has two levels (voiced, voiceless), manner two levels (fricative, occlusive), and place three levels (front, medium, back). In this paper, we assume that /ʒ/ and /ʃ/ are perceived as having a back place of articulation. A phonetic classification from these three feature dimensions allows for 2*2*3=12 classes, and there is no redundancy of transmitted information for the identification task. For example, if the stimulus is /f/, and if there is an error for manner identification only (/f/ is a fricative), then the response is /p/ (which is an unvoiced front occlusive).

[1] This work is supported by the EEC project RESPITE

feature	p	b	f	v	t	d	s	z	k	g	ʃ	ʒ
place	1	1	1	1	2	2	2	2	3	3	3	3
manner	o	o	f	f	o	o	f	f	o	o	f	f
voicing	-	+	-	+	-	+	-	+	-	+	-	+

Table 1: Coding of the articulatory features. Place 1: front, 2: medium, 3: back. Manner o: occlusive, f: fricative. Voicing +: voiced, -: voiceless.

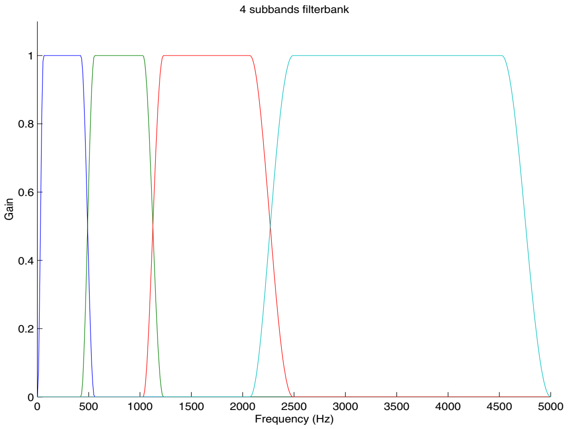


Figure 1: Filterbank design. The four quasi-rectangular filters have their center frequency at [267, 815, 1705, 3521] Hz.

One interest of the SRS in four subbands is the great simplification of the analysis of the relationship existing between the signal pattern and the phonetic dimensions: a clear spectral localisation of the feature dimension encoding appears at the signal level. A display of the four envelopes of each consonant class shows that the first subband (having center frequency at 267 Hz) carries the voicing/voiceless contrast whereas the third and fourth subbands carry the fricative/occlusive contrast, as well as some visible temporal events related to the residual place encoding.

2.3 Protocol of the perceptual experiment

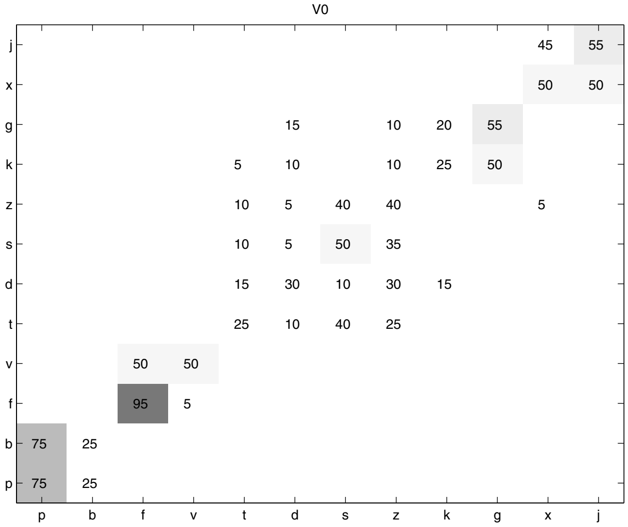
There were three experimental sessions named 'audio-visual complementarity', 'various combinations' and 'noisy condition' (these names become clear in the following). The same five people of ICP participated in the first and in the second session and four people in the last one. The main part of the results of these three sessions are presented here (other parts, not discussed here, tested the fullband SRS). This includes three blocks of AO presentation (10Hz, 500Hz and 10Hz+noise) and one block of VO. The VO stimuli were presented randomly mixed with AV ones during the first session and, similarly, the AO stimuli were presented during the two last sessions. For AV presentations, the same 48 video sequences were dubbed with SRS in four subbands for having coherent and incoherent /audio, visual/ pairs. In the later case, we applied a re-alignment of the audio track relatively to the video reference. The protocol of response was a 12 alternative forced-choice task without feedback. The subjects were informed to respond the class of the consonant they *heard* in the aCaCa stimulus even when they saw something they felt was incoherent (hence, excepted for VO stimuli). The subjects were not informed about the precise content of the experiment. In all sessions, incoherent /audio, visual/ pairs were presented, together with coherent ones. In other words, the different stimuli were randomised in order to have no strong prior information (this is an important point to fulfill for making the prediction of AV responses).

3. RESULTS

The responses were pooled across subjects and, for five subjects and four different records, there are 5*4=20 responses per stimulus type (inputs). These responses were compiled into confusion matrices and expressed in rounded percentages (easily converted in probabilities). For readability, figures 2-6 representing these confusion matrices are organised similarly: row input and column output. The output class consonant labels[2] are ordered for having the front to back place of articulation from left to right and the voicing confusions in the neighbourhood. For AV stimuli, the audio (x) - visual (y) input pairs are labelled xy.

3.1 Video and audio only

In the VO condition (Fig. 2), we retrieve confusions according to the four visemes classically described in the literature: two fronts {pb}, {fv}, a central {tdszkg} and a protrusion viseme {ʃʒ}. We remark that {kg} does not form a viseme, and that we have a lack of confusions between the stimuli /t/,/d/ and the responses /k/,/g/ respectively. In the AO confusion matrices (Fig. 3), the main characteristic of SRS in 4 subbands is observed: the place of articulation is degraded, and this corresponds with the bi-diagonal structure of the matrices. The place of articulation is more or less shifted backward for front and medium consonants: from front to medium and from medium to back. Moreover, for /f/ and /v/, there is also a shift from front to back. Other sparse confusions concern voicing and manner. There is no great difference between confusion matrices for the two conditions of filtering of the envelopes, 10Hz and 500Hz, and in the following, only the 10 Hz condition is figured.



VO

	p	b	f	v	t	d	s	z	k	g	x	j
j											45	55
x											50	50
g						15		10	20	55		
k					5	10		10	25	50		
z					10	5	40	40			5	
s					10	5	50	35				
d					15	30	10	30	15			
t					25	10	40	25				
v			50	50								
f			95	5								
b	75	25										
p	75	25										

Figure 2: VO confusion matrix.

[2] For a technical reason, /ʃ/ and /ʒ/ are noted /x/ and /j/ resp. in all Figures

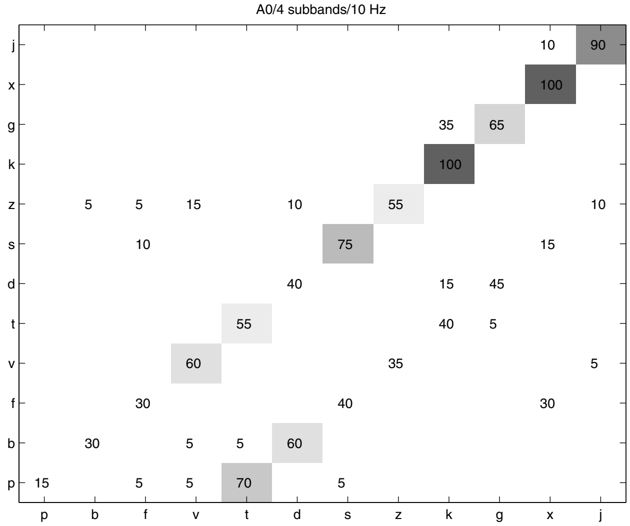


Figure 3: AO for SRS at 10 Hz (500Hz is not shown).

3.2 Coherent and incoherent audio-visual stimuli

The acoustic representation of the place of articulation (formant transition, release burst, aspiration, pole of the fricative noise) is degraded for moderate additive noise levels and in this condition, the audio confusions which are already present in the clean signal (e.g., /b/ vs. /d/ confusion) are reinforced. This weakness of the acoustical place encoding is the main motivation for lipreading and for engaging the property of audio-visual complementarity.

Furthermore, Sekiyama and Tohkura [11] have shown that the degradation of intelligibility is correlated with a more pronouced McGurk effect. This correlation means implicitly that the degradation of the place encoding induces the McGurk effect. As we have shown in Fig. 3, the use of SRS is a good method for having a well-controlled degradation of the place encoding. One advantage is that this kind of degradation is easier to understand at the signal level than the complex masking of the speech by white noise. The audiovisual SRSs at 10Hz and 500Hz have been tested during the 'audiovisual complementarity' session.

In Fig. 4, we see the AV confusion matrix (at 10Hz only) for coherent audio-visual pairs (top of the figure) and incoherent pairs (4 bottom raws of the figure). In the coherent section, there is only one spot of confusion between /tt/ and /k/, /dd/ and /g/. In the incoherent section, we observe a high amount of McGurk effect for the classical pairs /pk/ and /bg/. Remarkably, there is no front response and the normal McGurk effect is shifted backward (we also observe some back responses /k/ and /g/). Let's remark that the AO responses were already biased backward. So, we conclude that the McGurk mechanism produces an *additional* backward shift. This is probably the same mechanism, which is involved in the induction of the McGurk effect by noise, observed by [11].

We also introduce new McGurk combinations based on the fricative pairs /fʃ/ and /vʒ/. In Fig. 4, we observe the quasi-complete dominance of the video responses /ʃ/ and /ʒ/. At first, this can be explained in light of the preceding remark: in the AO condition, /f/ and /v/ are sometimes shifted to /ʃ/ and /ʒ/, so the backward bias is already strong and an additive shift leads to a complete attraction.

Another explanation is that {ʃʒ} form a specific viseme, and consequently, we have a simple dominance of the confident video over the ambiguous audio signal. To understand better this unexpected situation, we also take in consideration the production mechanism: The consequence of the protrusion is to lengthen the vocal tract, so there is a strong association between the protrusion and the acoustic characteristic of /ʃ/ and /ʒ/, which is normally represented by a low frequency pole (Badin, [1]). In SRS, the acoustic representation is greatly degraded; this is attested by the strong backward bias of the audio; so, this association is engaged, and this produces the visual dominance.

Furthermore, because the protrusion gesture could be interpreted as a normal back (velar) gesture, leading to an additive McGurk bias as we proposed as a first explanation. In the two cases, the goal of the gesture is to lengthen the vocal tract. This is the aim of the 'various combinations' section to test this hypothesis. We built four 'cross-manner' pairs /pʃ/,/bʒ/,/tʃ/,/dʒ/ in order to test if the protrusion gesture is interpreted as a back gesture leading to a McGurk effect because this is visible and has an additive effect. In this session, we also tested the original clean speech together with the SRS. These 'cross-manner' pairs are mixed together with coherent, as well as with incoherent pairs, which are not classical McGurk combinations.

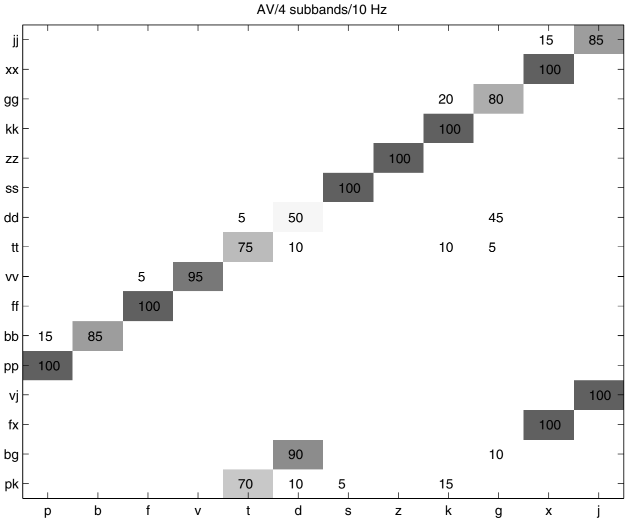
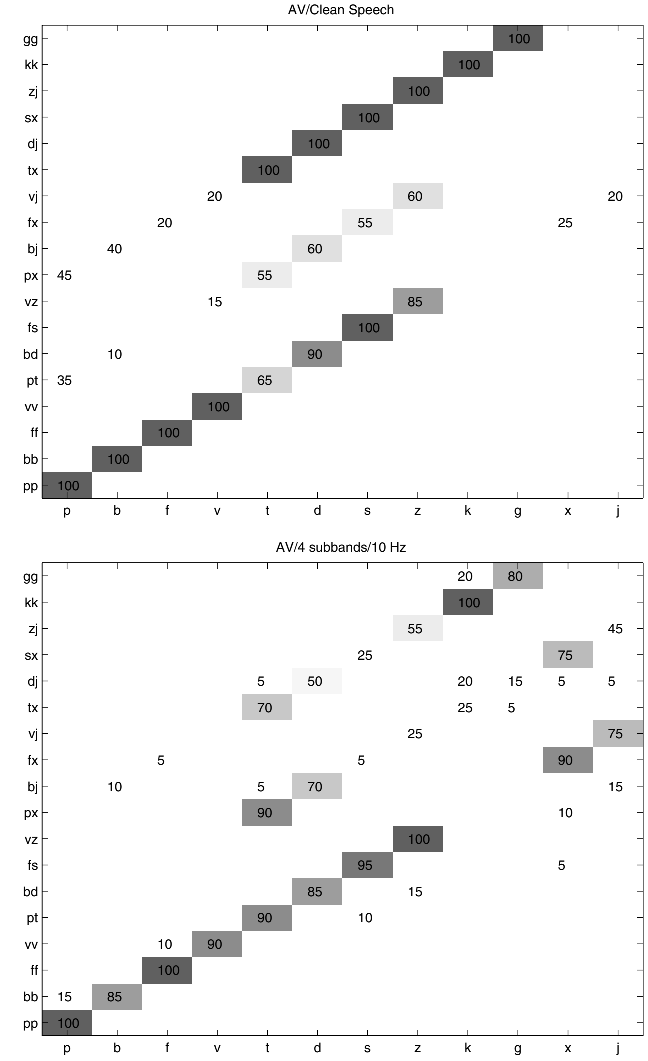


Figure 4: AV confusion matrix of the 'audio-visual complementarity' session (SRS at 10 Hz, 500 Hz not shown). The four incoherent (McGurk) combinations are placed in the bottom rows.

In Fig. 5 (top), we observe with clean speech a clear McGurk effect for /pʃ/ and /bʒ/ but smaller than for /pt/ and /bd/. With clean speech, there is no effect for /tʃ/ and /dʒ/, but this is the same as for /tk/ and /dg/ (this is not tested, but assumed) as well as for /sʃ/ and /zʒ/. For SRS speech (Fig. 5, bottom), the McGurk effect is observed in the four cases /pʃ/,/bʒ/,/tʃ/,/dʒ/ and this is stronger for the two first pairs /pʃ/,/bʒ/. For the /tʃ/,/dʒ/ pairs there is small effect specific of SRS (we name "middle McGurk"), as well as the strong effect for /sʃ/ and /zʒ/. All of this happens as if the protrusion is interpreted as a back gesture, and this fulfils the 'additive bias hypothesis'. The other pairs used in this experiment will be used for testing a model, which allows a quantification of this hypothesis.



AV/Clean Speech

	p	b	f	v	t	d	s	z	k	g	x	j
gg										100		
kk									100			
zj								100				
sx							100					
dj						100						
tx					100							
vj				20				60				20
fx			20				55				25	
bj		40				60						
px	45				55							
vz				15				85				
fs							100					
bd		10				90						
pt	35				65							
vv				100								
ff			100									
bb		100										
pp	100											

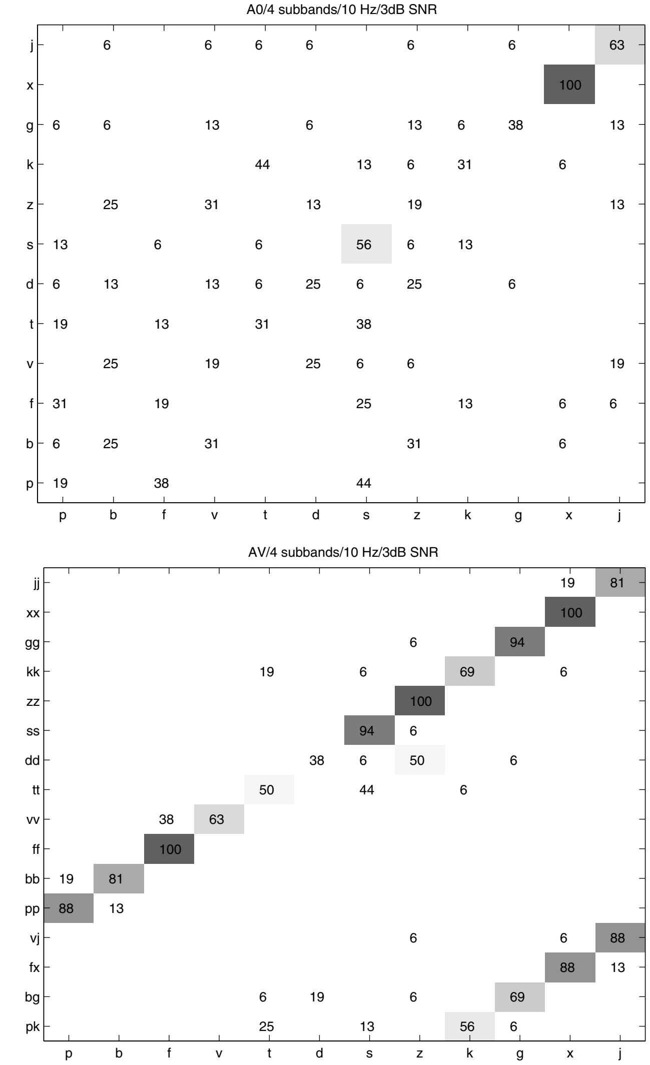
AV/4 subbands/10 Hz

	p	b	f	v	t	d	s	z	k	g	x	j
gg									20	80		
kk									100			
zj								55				45
sx							25				75	
dj					5	50			20	15	5	5
tx					70				25	5		
vj								25				75
fx			5				5				90	
bj		10			5	70						15
px					90						10	
vz								100				
fs							95				5	
bd						85		15				
pt					90		10					
vv			10	90								
ff			100									
bb	15	85										
pp	100											

Figure 5: AV confusion matrices of the 'various combinations' session. Top: clean audio speech; bottom: SRS at 10Hz (500 Hz is not shown).

3.3 Noisy condition

In our set of 12 consonants, the fricative/occlusive contrast is the only manner distinction. A previous experiment with audio SRS has shown that this contrast is suppressed when low frequency (< 8Hz) modulated white noise is added at 6dB (Grosgeorges et al. [5]). On the contrary, the transmission of voicing and place was relatively less degraded. So, this type of additive AM noise is rather selective for removing the audio manner characteristics. Since the video signal well transmits the manner feature ({pb}, {fv} and {ʃʒ} have distinct visemes), this allows to test the audio-visual complementarity and the McGurk effect when manner transmission is degraded. In the noisy condition, we add to SRS at 10Hz a similar modulated noise at 3dB SNR. In comparison with Fig. 3 (SRS at 10Hz without noise), where we observe a bi-diagonal distribution, Fig. 6 (top) shows a great dispersion of the confusions. An attentive inspection reveals that the main cause is the degradation of the manner recognition, as expected. This dispersion is greatly reduced in the AV condition (Fig. 6, bottom), and this shows that the AV complementarity plays a great role in manner transmission: The confusions are reduced thanks to the visemic transmission of the manner.



A0/4 subbands/10 Hz/3dB SNR

	p	b	f	v	t	d	s	z	k	g	x	j
j		6		6	6	6		6		6		63
x											100	
g	6	6		13		6		13	6	38		13
k					44		13	6	31		6	
z		25		31		13		19				13
s	13		6		6		56	6	13			
d	6	13		13	6	25	6	25		6		
t	19		13		31		38					
v		25		19		25	6	6				19
f	31		19				25		13		6	6
b	6	25		31				31			6	
p	19		38				44					

AV/4 subbands/10 Hz/3dB SNR

	p	b	f	v	t	d	s	z	k	g	x	j
jj											19	81
xx											100	
gg								6		94		
kk					19		6		69		6	
zz								100				
ss							94	6				
dd						38	6	50		6		
tt					50		44		6			
vv			38	63								
ff			100									
bb	19	81										
pp	88	13										
vj								6			6	88
fx											88	13
bg					6	19		6		69		
pk					25		13		56	6		

Figure 6: AO (top) and AV (bottom) confusion matrices of the 'noisy condition' session. The four incoherent combinations are placed in the bottom rows.

The main residual confusions are between /tt/ and /s/, /dd/ and /z/. Interestingly, these are manner confusions which cannot be removed by the central viseme and the inverse confusions, /ss/ and /t/, /zz/ and /d/, are not observed. For /pk/ and /bg/ combinations, a McGurk effect is observed which is dispersed among the central viseme, but we remark that /k/ and /g/ form a majority of responses. This point will be discussed later. For the two other combinations /fʃ/ and /vʒ/ the visual dominance of /ʃ/ and /ʒ/ is observed for SRS at 10Hz without noise (Fig. 4).

4. TWO PREDICTION MODELS

4.1 The Bayesian approach applied to automatic speech recognition

A unified point of view is presented by Massaro and Stork [8] for modelling the process of perceptual AV fusion (complementarity and McGurk) and then transposed to the same model in automatic recognition. The authors assume that audio and visual estimates are combined in a late fusion architecture. Furthermore, the assumption of class conditional independence allows for applying the Bayes rule in order to fuse the separate audio and video estimates of the posteriors:

$$P(c \mid A,V) = P(c \mid A).P(c \mid V).\varepsilon \qquad \text{(Eq. 1)}$$

where ε is the normalisation coefficient.

In the literature, many simulations show that this statistical modelling approach has good results for automatic audio-visual speech recognition. The late fusion architecture has been adopted as a standard in this field. One reason is the Bayes formulation of AV fusion is compatible with the state of the art of speech recognition systems.

In a companion paper (Heckmann et al. [6]), we show an implementation of this model, carried out in the context and with the tools of multi-stream recognition. The aim of multistream recognition is to exploit the redundancy of multiple or partial representations (n≥2) of the input in order to compensate when some streams are missed or when their content is noisy and distorted. In our application to audio-visual recognition, the estimates of the posteriors are delivered frame by frame by an ANN stage and the task is continuous robust recognition of words in a small vocabulary. At first, we apply a known improvement of Bayes rule, which consists in weighting the posteriors according to an estimate of their reliability:

$$P(c \mid A,V) = P(c \mid A)^{a}.P(c \mid V)^{(1-a)}.\varepsilon \qquad \text{(Eq. 2)}$$

where $a \in [0, 1]$.

Furthermore, we introduce the prior values (see [6]). The knowledge of priors allows an improvement of continuous word recognition in noise because the phoneme frequency is not uniform and because the silence state (also having a prior value) has to be detected whatever the noise level. We get a good AV synergy, which fulfils the 'product of errors' criterion at the word level: for each level of SNR, the AV error rate is the product of AO and VO error rates. Remarkably, our model reaches this bound criterion of multistream recognition. Streams' estimates are independent and apparently, errors occur independently in each channel. In this condition, errors in the audio channel are well compensated by the video and vice versa, since the channel which is reliable at a given time is apparently perfectly detected at the word level. The only residual errors, which cannot be recovered, are those occurring in the two channels at the same time.

4.2 Modelling of human performances

But performances are expected to overcome this boundary for audio-visual recognition because, thanks to the audio-visual complementarity, the errors are precisely not independent, but anti-correlated at the phonetic feature level; e.g. front/medium confusions are present in the audio only whereas medium/back errors occur in the video only. Most of the published results about perceptual experiments, showing an AV %rec. vs. SNR, overcome this boundary and human outperforms the 'product of error' criterion. So, for machines, a progress is possible from a better exploitation of the natural audio-visual complementarity.

Hence, [8] argue that the Bayes rule is satisfactory for capturing the effect of audio-visual complementarity at the phonetic class level. They predict the AV confusion matrix from AO and VO confusion matrices, using the (Eq. 1), for coherent AV stimuli as well as for McGurk combinations. For SRS, a rapid inspection of VO (Fig. 2) and of the three related pairs of AO (Fig. 3 and 6) and coherent AV (Fig. 4 and 6) matrices leads to the conclusion that AV is rather similar to the product of VO and AO. However a detailed analysis reveals the limits of this method. To obtain good predictions of the five AV confusion matrices (3*16+2*18 stimuli) from their related AO and VO, we apply four corrections at different levels, listed by order of importance:

- For the 'cross-manner' combinations /pʃ/, /tʃ/, /bʒ/, /dʒ/, the back fricative is substituted by the resp. back occlusive /k/ and /g/ because, otherwise, the response is not predictable.
- The VO confusion matrix is regularised by assigning 50% to the diagonal and by a symmetric repartition of the other 50% off diagonal within the 4 sub-matrices of the 4 visemes. This is required for the central viseme sub-matrix (see Fig. 2), in which we have to create symmetric confusions not observed in our experiment.
- When the result is indeterminate, because this is based on a small percentage, which is probably not represented in the AO matrix because the experiment is too small, this is simply added (only 1 stimulus is concerned).
- The eq. 2 is used with a=0.7 for having better quantitative results.

After these corrections, the Bayesian approach combined with the late fusion architecture allows to describe the overall pattern of human responses well. But there are some limitations and in many cases, the predicted confusions are concentrated, and depend on a small percentage of response existing in AO and VO matrices. We conclude these problems are due to the multiplicative nature of the prediction, which is only based on the overlap of tails of AO and VO class consonant confusions. Then, the main modifications we apply consist in creating a non-existing overlap of the two distributions. Particularly, the backward shift observed for SRS 'cross-manner' pairs /pʃ/, /tʃ/, /bʒ/, /dʒ/ cannot be explained by this model because there is no overlap between the VO output class distribution of the protrusion viseme and the observed distribution of AO responses. Similarly, the majority of /k/ and /g/ responses to the McGurk combinations /pk/ and /bg/ in the SRS noisy condition (Fig. 6 bottom) cannot be predicted. In this condition, there is a lack of transmission of manner by the audio signal (Fig. 6 top). Consequently, in AO, dominant responses to /p/ and /b/ are /s/ and /z/, and there is no response for /k/ and /g/ (this is not evident these responses will appear in a larger experiment). Then, the predicted AV responses to /pk/ and /bg/ are resp. /s/ and /z/ and the Bayes rule fails to predict /k/ and /g/.

To explain the observations, we propose a model (named AFC, Articulatory Feature Coding) which is based on averaging of the distribution of responses, as in a similar approach promoted by Braida et al. [2] who used the *centers* of distributions instead of the tails. The main point of our model is the decomposition of the recognition process into streams specific for each articulatory feature. The idea is to have a better exploitation of the distribution of confusion along a given feature dimension. Moreover, this allows expressing specific rules of audio-visual fusion for each of these articulatory features.

At first, we suppose that the identification of the phonetic class is preceded by a separate evaluation of each feature (voicing, manner and place), independently for audio and video. These evaluations are fused according to a rule, and we propose that, at least for place coding, this rule be based on averaging. Then, to analyse the confusion matrices, for each stimulus, the averaging process consists of converting the response classes in articulatory feature levels, and then weighting by the class frequency of response. For the place of articulation, encoded $X \in \{\text{front,medium,back}\} = \{1,2,3\}$, the evaluation arising from confusion matrices is the average place code of the responses for a given stimulus s (M=A, V, or AV):

$$F_M(s) = \sum_c P(c \mid M).X(c) \qquad \text{(Eq. 3)}$$

The prediction rule of the AFC model is applied to AO and VO confusion matrices. This is the following:

$$\text{If } F_V(s) = 1 \quad \text{then} \quad F_{AV}(s) = 1$$
$$\text{else} \qquad \text{(Eq. 4)}$$
$$F_{AV}(s) = a\, F_A(s) + (1-a)\, F_V(s)$$

with a=0.3

For the place feature, this means that the front video articulation is dominant, and that, otherwise, the evaluation is an average between audio and video estimates. In Figure 7, the average place of response, Fav, predicted with the AFC rule vs. this observed from the AV matrices (Eq. 3), is plotted for each stimulus (O) of the five experimental blocks. In the same figure, the Fav obtained with confusion matrices predicted with the corrected Bayes rule are plotted (*) for comparison. The correlation coefficient between observed and predicted values is high for the two models (resp. 0.965 for AFC and 0.972 for Bayes). The Bayes rule has a small tendency to overestimate (i.e., this is too backward), and the AFC to underestimate the place of articulation. The main difference is the absence of any supplementary assumption for the AFC model.

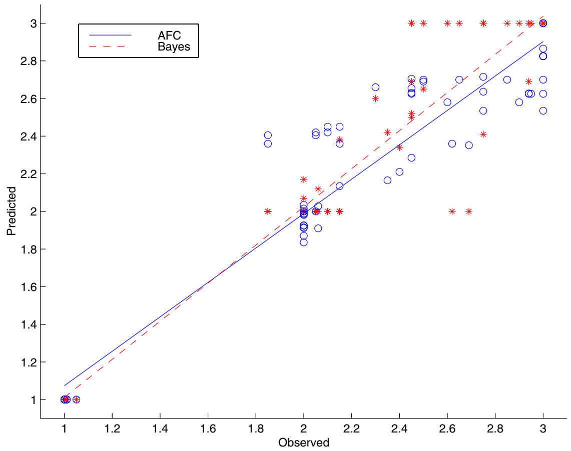


Figure 7: Comparison between the prediction of place from VO and AO (within 1: front and 3: back) using the corrected Bayes rule (*) and with the AFC model (O) vs. the observation calculated with AV confusion matrices.

Other experiments are necessary for evaluating the pertinence of a similar AFC rule for fusion of the manner (e.g., experiments with varying SNR). From the present experiment, what we know about manner transmission is the following: **(1)** In the noisy condition allowing a suppression of the manner transmission in the audio signal, this is mainly determined by the video (Fig. 6). **(2)** There are few false manner responses observed in the AO without noise (Fig. 3), so this is also well transmitted by the audio. **(3)** These few errors are well filtered out by the video in the AV condition (Fig. 4). **(4)** With the cross-manner stimuli, the manner is not inherited from the video and the audio is dominant (Fig. 5). The Bayes rule is valid for explaining the first three points.

5. CONCLUSION

These results obtained with audio-visual SRS suggest an alternative method to the standard 'one way' phoneme classification used in automatic speech recognition. We propose to evaluate independently the phonetic features, which ground the classification process in separate streams each specialised for an articulatory feature. It was admitted since Miller and Nicely (1955) that the consonant classification process relied on articulatory features, but the identification of specialised processes and the localisation of their support remained problematic for clean speech and noisy speech (Grant and Walden, [4]). The use of SRS allows a great simplification at the signal level, and this clarifies the problem of the representation of the articulatory feature dimensions, as well as of their audio-visual fusion. The AFC model is an alternative way of evaluating the place of articulation instead of the phoneme class. The perception of 'cross-manner' McGurk combinations as /pʃ/, /tʃ/, /bʒ/, /dʒ/ can be explained with this model, assuming that /ʒ/ and /ʃ/ are perceived as having a back place of articulation. This initial assumption is consistent with our results. We conclude that audio-visual SRS recognition is a powerful paradigm for better understanding of the audiovisual fusion process.

Acknowledgements: This work is a part of the RESPITE project. Thanks to Pierre Badin for useful comments, to Angélique Grosgeorges and Stéphanie Brosda for their technical aid for recording the audio-visual database, and to the ICP people who participated in the experiments.

References

[1] Badin, P. (1989) Acoustics of voiceless fricatives: production theory and data, STL-QPSR 3/1989, pp. 33-55.

[2] Braida, L.D., Sekiyama, K. & Dix, A. K. (1998) Integration of audiovisually compatible and incompatible consonants in identification experiments, proc. of AVSP'98, pp. 49-54, Terrigal.

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

[4] Grant, K. W. & Walden, B.E. (1996) Evaluating the articulation index for auditory-visual consonant recognition, JASA, 100 (4), 2415-2424.

[5] Grosgeorges, A., Berthommier, F. & Lorenzi, C. (2000) Effect of masking of spectrally reduced speech with modulated noise, JASA, 140th meeting, 108 (5), 2602.

[6] Heckmann, M., Berthommier, F. & Kroschel, K. (2001) A hybrid ANN/HMM audio-visual speech recognition system, this volume.

[7] Horii, Y., House, A. S. & Hughes, G. W. (1971) A masking noise with speech-envelope characteristics for studying intelligibility, JASA, 49, 1849-1856.

[8] Massaro, D.W. & Stork, D. G. (1998) Speech recognition and sensory integration, American Scientist, 86, 236-244.

[9] McGurk H. & MacDonald, J. (1976) Hearing lips and seeing voices, Nature, 264, 746-748.

[10] Miller, G.A. & Nicely, P. E. (1955) An analysis of perceptual confusions among some English consonants, JASA, 27, 338-352.

[11] Sekiyama, K. & Tohkura, Y. (1991) McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing japanese syllables of high auditory intelligibility, JASA, 90, 1797-1805.

[12] Shannon, R. V., Zeng, F-G., Kamath, V., Wygonsky, J. & Ekelid, M. (1995) Speech recognition with primarily temporal cues, Science, 270, 303-304.

[13] Van Tassel, D. J., Soli, S. D., Kirby, V. M. & Widin, G. P. (1987) Speech waveform envelope cues for consonant recognition, JASA, 82 (4), 1152-1161.