Joint Matrix Quantization of Face Parameters and LPC Coefficients for Low Bit Rate Audiovisual Speech Coding

Laurent Girin

***Abstract*—A key problem for videophony, that is telephony including the processing of images of the speaker's face in addition to acoustic speech, concerns signal compression for transmission. In such systems, audio and video compression are separately achieved by using both audio and video coders. In this paper, an audio-visual approach to this problem is considered, since we claim that the fundamental property of coherence (redundancy) between the two modalities of speech should be exploited by coding systems. We consider the framework of parametric analysis, modeling and synthesis of talking faces, which allows efficient representation of video information. Thus, we propose to jointly encode several face parameters, namely lip shape geometric descriptors, together with sets of audio coefficients, namely quite usual LPC parameters. The definition of an audiovisual distance between vectors of concatenated audio and video parameters allows to generate audiovisual single stage vector and matrix quantizers by using the generalized Lloyd algorithm. Calculation of video and audio mean distortion measures shows a significant gain in quantization accuracy and/or resolution compared to separate video and audio quantization. An alternative sub-optimal tree-like structure for audiovisual joint coding is also tested and yields interesting results while decreasing the computational complexity of the quantization process.**

***Index Terms*—Audiovisual, lip parameters, low-bit-rate speech coding, LPC parameters, matrix quantization, speech processing.**

I. Introduction

SPEECH is both an acoustic and a visual signal, that is sequences of different sounds together with visible movements of the speaker's face. This bimodal nature of speech is now acknowledged as a basic characteristic, both for understanding speech perception [1] and for developing tools for human-human and human-machine communication [2]. The major characteristic of audio-visual speech is that there is some dependence between the audio and visual signals [3], since they are both consequences of the articulatory gestures: e.g., spread lips may be associated in French with the sound of an [i] or a [ti] but not an [y] or a [by], while open lips are compatible with almost nothing but the sound of an [a].

In the field of speech technology, there is an increasing number of work dealing with how to exploit this coherence between audio and visual speech signals to improve the performances of classical speech processing systems. As a major example, automatic audiovisual speech recognition systems have recently received considerable interest [4], notably for their ability in recognizing speech in adverse conditions. In this latter case, video speech, often restricted to geometric lip shape descriptors, can be considered as supplying complementary information when audio information is degraded. In a recent study [5], we demonstrated the technical feasibility of audio-visual speech enhancement, that is, the enhancement of noisy speech sounds, using a video input. In this study, denoising filters were estimated from noisy speech plus speaker's lip shape information.

In this paper, we consider the problem of audio-visual speech coding, that is the compression of both audio and video information to be transmitted or stocked in a telecommunications system. Indeed, the processing and transmission of the speaker's face image has become a major challenge for future telecommunications systems, as illustrated by the recent development of videoconferencing systems or multi-modal human-computer interfaces. But if we increase the amount of information (channels) to be transmitted, we must deal with efficient techniques to compress this information before transmission. Thus, accurate audio and video coding techniques have been separately developed in order to minimize the quantity of binary streams to be transmitted. These techniques aim to reduce the redundancy of each audio or video signal. As a major example, predictive coding is widely used to reduce the intra- or inter-frame correlation between consecutive (time or space) samples of signals. In this context, the natural cross-modal coherence between audio and visual speech can be seen as a form of redundancy, which should also be exploited (ideally eliminated) by any coding algorithm. As a matter of fact, the objective of this study is to demonstrate that a joint coding scheme of audio and visual speech parameters can lead to better performances than separated coding algorithms.

The context of this study is talking faces and a typical and major application is videophony, a telephone enhanced with the image of the speaker's face. Now, if we consider any given video sequence, typical bit rates are around several hundreds of kbits/s. As telephone speech can be efficiently coded at around 5 kbits/s (e.g., using CELP techniques), the interest of jointly coding sound and video can be questioned by this asymmetry. However, this difference can be highly reduced by exploiting the high specificity of talking faces images in videophony context: slow changes of the (less important) background and upper part of the face, and rapid changes of highly informative local zones, mainly the mouth and the jaw and to a lesser extend the

Manuscript received April 24, 2002; revised October 16, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Vary.

The author is with the Institut de la Communication Parlée, INPG/Université Stendhal/CNRS, 38031 Grenoble, France (e-mail: girin@icp.inpg.fr).

Digital Object Identifier 10.1109/TSA.2003.822626

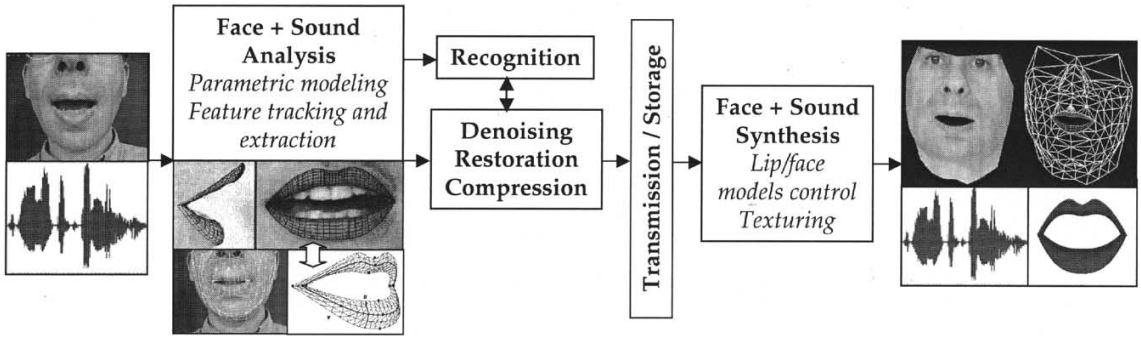

Fig. 1. Basic schema of an audiovisual speech processing/transmission system based on analysis, modeling and synthesis of audiovisual speech signals.

cheeks and eventually the eyes and eyebrows. Although they extensively use differential predictive coding, current videophone algorithms still do not optimize the use of this specificity as they function around 100 kbits/s for the image. Thus, an additional step forward to better exploit the specificity of talking faces in coding applications may be 2-D/3-D face modeling before and after transmission. The principle is to segment the speaker's face into regions that basically represent the different organs of the face, and then to parameterize these regions with a small number of parameters describing the local geometry and texture. The accuracy and the refreshment of the parameters must ideally be adapted to the degree of information contained and the speed of changes. For example, the mouth should be described with care. This naturally leads to a parametric low bit rate coding of talking faces: useful varying information is concentrated in a small number of parameters. At the receiver, the speaker's face is synthesized by applying the transmitted (quantized) parameters over a generic 2-D or 3-D face model. The parameters may eventually be separated into parameters that encode the (fixed and neutral) configuration of the face (e.g., the Face Definition Parameters of MPEG-4 [6], [7]) and that are to be transmitted only once in a given communication situation (if not already available at the decoder in a speaker database), and parameters that encode the movements of the face (e.g., the MPEG-4 Facial Animation Parameters) and that should be transmitted with an adapted rate.

The problem of face analysis for parameter extraction as well as face 2-D/3-D modeling and synthesis is of growing interest in the speech and image research community (for a recent review of face synthesis, see [8]) and should benefit to audio-visual speech coding. For example, a 3-D articulatory-based speaker-specific talking head model has recently been developed at the ICP [9]. It is controlled by 6 parameters of articulatory nature (e.g., two of them are closely related to the movement of the jaw and three are closely related to the round and open movements of the lips). These parameters were derived from data analysis of around 200 face fleshpoint measurements and they explain 96% of the variance of the data. Therefore, they can be used to efficiently control the movement of the fleshpoints, and thus of the entire face by using interpolation techniques. Since such efficient parametric face modeling may allow low bit rate talking face video coding, we can now assume that joint coding of facial and audio parameters is realistic from a bit rate criterion. Summarizing, we illustrate in Fig. 1 the principle of an audio-visual speech transmission system (videophone), exploiting both parametric face description and joint audio-visual coding, as well as other possible applications that were discussed above (e.g., enhancement and recognition). Note that in this study, only geometric lip shape parameters were considered for the video data. These parameters, the justification of their choice, and their acquisition process are described in more details in Section II-A (the audio parameters to be jointly encoded are also described in this section).

To achieve joint quantization of lip shape and audio parameters, we had to use a technique allowing: (1) quantization of a set of parameters, (2) quantization of parameters different in nature. We used vector quantization (VQ) and matrix quantization (MQ). VQ not only allows by definition to deal with condition (1) but is known to be an efficient method for exploiting the correlation between data vector components. MQ was used as a generalization of VQ to remove inter-frame correlation of consecutive vectors, as we can expect the different sets of audio and visual parameters to be correlated in time. The condition (2) was resolved by defining an audio-visual distance in the heart of the VQ/MQ process. In Section II, we describe in more details the main VQ/MQ schema used in this study, including a brief review of the "classical" quantization of the audio parameters that were used, the LSPs. We then define the audio-visual distance and describe the training process for the VQ/MQ codebooks. Moreover, at the end of the section, we propose an alternative so-called "classified-VQ/MQ" structure (C-VQ/C-MQ), which is also able to benefit from the audio-visual coherence while presenting different advantages (and drawbacks). In Section III, results are presented. We achieved a complete distortion-rate quantitative evaluation for both joint audiovisual coding and separated audio and visual coding used as references, leading to a quantitative estimation of the gain provided by joint coding over separated coding.

It is important to note here that the idea of a joint cross-modal speech coding process was originally proposed by Rao and Chen [10], who largely explored the extended field of audio-visual speech processing [11]. In [10], they proposed a predictive coding system for geometric labial templates where the prediction was made from acoustic parameters by using an audio-visual probability distribution of Gaussian mixtures. Unfortunately, the corpus that was used both for tuning the mixture parameters and testing the system was limited to four vowels, thus only allowing to grossly validate the feasibility of

the approach. In this paper, we deal with an extended corpus of multi-speaker fluent speech (see Section III-A) and an extensive use of vector/matrix quantization and LSP representation of audio parameters, which are techniques that are largely used in state-of-the-art speech coders. Therefore, we tend to get close to the "true" coders and we provide a complete quantitative evaluation of the joint audio-visual coding scheme for real implementations.

II. Joint Audio and Video Parameters Vector/Matrix Quantization

A. Audio-Visual Data

As briefly mentioned before, the visual data that was considered in this preliminary study are basic geometric lip contour parameters. Four parameters are considered: internal and external width, (respectively denoted *LIW* and *LEW*,) and internal and external height, (respectively denoted *LIH* and *LEH*.) Such parameters can be extracted every 20 ms[1] by different face processing systems developed at the ICP [12]. The current system works with blue make-up for the lips using the Chroma-Key technique, but recent studies have shown encouraging results for the extraction of similar parameters on natural lips [13]. One version of the system is portable (camera on headphones) and resolves most problems relative to acquisition conditions. Other version of the system allows the acquisition of other parameters, e.g., profile parameters like upper and lower protrusion and position of the corner of the lips. However, we restricted this study to four facial lip parameters, and no additional articulatory parameters (e.g., as the ones used in the articulatory talking head model of [9]) were considered because of the following reasons. Easiness of acquisition comes first: the study presented here is greedy regarding the corpus that was necessary to use to provide significantly representative results (see Section III-A). Now, the analysis of fleshpoints and the extraction of articulatory parameters, which are based on an analysis-by-synthesis process, are currently computationally too fastidious to provide an extended corpus, and they require a speaker-specific model for each speaker. On the contrary, as mentioned above, the lip parameters can be efficiently and rapidly extracted with the Chroma-Key-based face processing system. Secondly, several studies have shown that the basic facial lip contour parameters contain most of the visual information, according to both intelligibility criterion [14], [15] and statistical analysis: the internal width and height represent 85% of the variance of the visual data used in [9] (to be compared with 96% obtained with the 6 articulatory parameters). As a consequence, if correctly quantized, these parameters can be used at the receiver to drive a 3-D lip model in an efficient manner, that is preserving a major part of the visual intelligibility, as shown in an early study [15]. Even further, the lip parameters can be used to estimate the articulatory control parameters of [9] via linear regression based transformations and vice-versa without major loss of information. Therefore, they can be used to control the complete articulatory 3-D face model.[2]

Note finally that the external width and height were also used in the current study because they were automatically extracted together with the internal parameters, and although they are generally redundant with them, they can provide quite useful additional information in certain cases (e.g., the external width contains a large part of the round information when lips are closed or almost closed). Adding these parameters was not a problem since the VQ/MQ is able to exploit/remove their redundancy with a very small additional CPU cost.

In what concerns the audio parameters, we considered the classical linear prediction coding (LPC) context, that is a "source-filter" scheme for low bit rate speech coding, where the "filter" part is achieved by applying a linear prediction model on the speech samples. LPC coding has been shown to be a very efficient technique for low bit rate coding of telephone-band speech, mainly because of its ability to exploit the intra-frame time correlation of speech samples. Thus, it is now employed in most of the voice coders for telephony which differentiate from each other by different bit rate and quality according to (1) the way the LPC "filter" parameters are quantized and (2) the way the "source" is encoded and eventually modeled. Now, if we consider the objective of our study, that is melting audio and video oro-facial parameters in the coding process, we can highly benefit from the LPC model by remarking that the "filter" part of the "source-filter" corresponds to the modeling of the vocal tract transfer function (thus the short-term spectral envelope of speech) while the "source" part correspond to the modeling of the excitation source of speech, that is vocal cord vibration and/or air flow. Now, we can generally expect the oro-facial parameters to be correlated with the vocal tract shape, thus with the filter parameters, and not (as much) with the excitation or source which is not visible. This is particularly true for the lip shape parameters that are used in this study: they represent in fact the most visible part of the vocal tract and thus the most important part of visual intelligibility. This is why in this study, we chose to jointly encode lip parameters together with the LPC filter parameters, and with the LPC filter only.

In most up-to-date coders, the LPC parameters are transformed before quantization to a more efficient representation that is more suitable for quantization: the Line Spectrum Pairs (LSP) which consist in (normalized) frequency values around the LPC poles arguments [17]–[19]. This representation offers suitable statistical properties and remarkable robustness to low bit rate quantization. Therefore the audio data to be jointly quantized with the lip parameters are sets of 10 LSP parameters, as 10 is a standard order for the filter of most LPC-based telephone band speech coders. The problem of extracting the LSPs from the sound and synchronizing them with the sets of video parameters is discussed in Section III-A.

[1]Measurements were made on PAL standard videos at 25 images/s and 2 interleaved frames per image: one parameter set is extracted on every separated frame, resulting in a 50 Hz sampling period.

[2]It was also shown in [9] and [16] that the articulatory parameters can be efficiently transcoded into MPEG-4 Facial Animation Parameters with similar transformations and vice-versa. Therefore, we can expect our lip parameters to be possibly "compliant" with this norm through transcoding routines, though this point is currently out of the scope of this study. Of course, future work may also involve direct use of MPEG-4 parameters for joint coding.

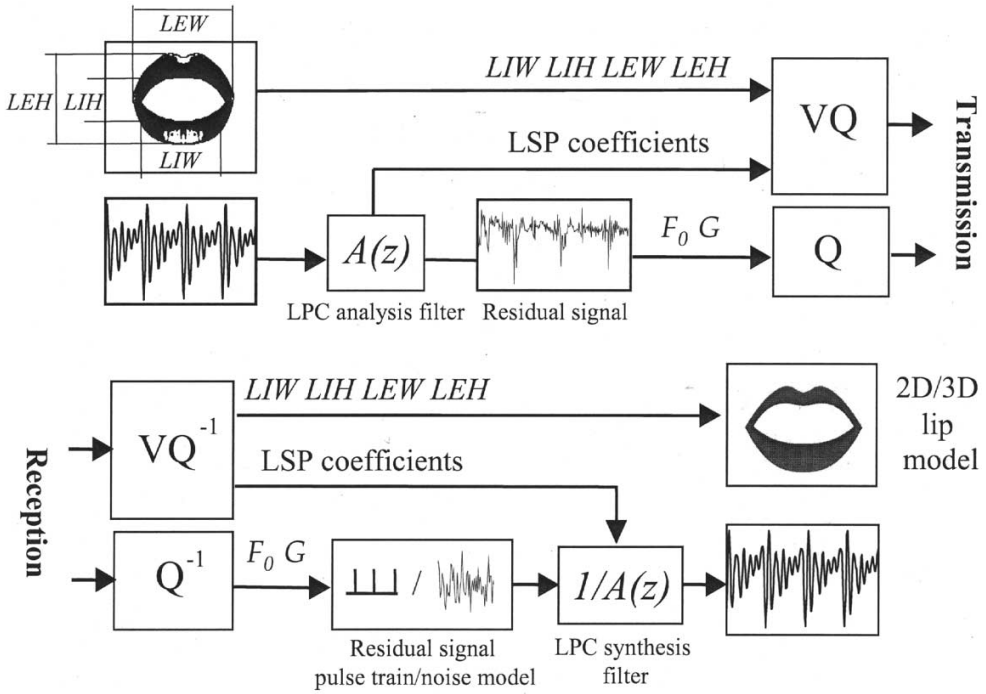

Fig. 2. Structure of an audiovisual speech codec based on audiovisual single stage vector quantization of lip and LSP parameters. Residual acoustic signal parameters (e.g., gain G and fundamental frequency F_0) are quantized and transmitted separately.

B. Brief Review of LSP Vector/Matrix Quantization

Vector quantization has been extensively used to efficiently quantize LSP coefficients [18], [19]. A basic single stage VQ was implemented at 20 bits/frame [20] (a frame corresponds here to a set of 10 LSPs) and offered high coding quality while basic scalar quantization of LSP parameters would typically involve around 40 bits/frame. However, in this resolution range VQ implementation is limited by codebook storage capacity, search complexity and training procedure. Thus different schemes have been propose to reduce complexity by breaking the quantization process in several steps. Split-VQ (SVQ), which consists of splitting the vectors into several sub-vectors for quantization have been proposed at 24 bits/frames and offered coding transparency[3] [18]. Multi-stage VQ (MS-VQ) consists in cascading several "low resolution" VQ blocks. The output of a block is a quantization error which is quantized by the next block, the signal being reconstructed by adding the outputs of the different blocks. Therefore, each successive block increases the quantization precision while the global complexity (in terms of codebook generation and search) is reduced compared to a single-stage VQ with the same overall bit rate. Indeed, the codebook size of each sub-block is highly reduced compared to the size of the "equivalent" single-stage codebook. Such schemes have been successfully implemented for LSP coding [21], [22].

Differential and/or adaptive VQ have been proposed to remove interframe correlation between consecutive sets of LSP coefficients [23] leading to high quantization accuracy at around 20 bits/frame. Matrix quantization (MQ) is a generalization of VQ which also allows to remove inter-frame correlation by jointly coding successive sets of LSPs in a simpler manner (these are gathered in a matrix and replaced by matrix codewords) but at the price of an increased delay and calculation cost [24]. Note that most of these schemes can be combined at different levels. For example, many different schemes proposed for VQ can be applied to MQ (see for example a Split MQ design in [25]).

[3]"Coding transparency" of LPC filter parameters means that speech sequences synthesized with the quantized and unquantized LPC parameters are perceptually undistinguishable

C. Joint Lip Parameters and LSP Single-Stage Vector/Matrix Quantization

In this study we want to demonstrate the advantage of jointly quantizing audio and video parameters in a joint quantizer block compared to audio and video separated quantization with two quantizer blocks both based on the same techniques than the global quantizer block. To show this and because fine tuning of sophisticated blocks was not our objective, we chose to use quite simple structures: single stage VQ/MQ have been mainly considered. Thus an audio-visual single stage VQ (AV-SS-VQ) is first proposed, where an audio set of LSP coefficients (corresponding to a frame of acoustic signal) is concatenated with a synchronous video set of lip parameters (see Fig. 2). This AV-SS-VQ has been implemented and compared with an audio-only single stage VQ (A-SS-VQ), that is "classical" VQ of LSP coefficients, together with a video-only single stage VQ (V-SS-VQ), that is VQ of lip parameters. Then, this scheme has been extended to MQ, leading to an audio-visual single stage MQ (AV-SS-MQ), audio-only single stage MQ (A-SS-MQ) and video-only single stage MQ (V-SS-MQ) respectively gathering successive sets of audio-visual, audio only and video only vectors.

The definition of a distortion measure is a major point in the design and application of a VQ/MQ. In this study, this point is especially crucial, since the processing of audio-visual vectors or matrices asks for a distortion measure that can take into account the difference in nature of the audio and visual components of the data. We chose to use an audio-visual distortion measure that is a linear combination of audio and video distortion measures:

$$D_{AV}\left(X,\hat{X}\right)=\sum_{i=1}^{M}\alpha D_V\left(V^i,\hat{V}^i\right)+(1-\alpha)\,D_A\left(A^i,\hat{A}^i\right) \tag{1}$$

where $V^i=(LIH^i,LEH^i,LIW^i,LEW^i)^t$ and $\hat{V}^i=(L\hat{I}H^i,L\hat{E}H^i,L\hat{I}W^i,L\hat{E}W^i)^t$ denote two column vectors of lip parameters (one may be data and the other a codeword), and $A^i=(LSP_1^i,LSP_2^i,\ldots,LSP_{10}^i)^t$ and $\hat{A}^i=(L\hat{S}P_1^i,L\hat{S}P_2^i,\ldots,L\hat{S}P_{10}^i)^t$ denote two column vectors of LSP parameters (idem). In this study, $D_V\left(V^i,\hat{V}^i\right)$ and $D_A\left(A^i,\hat{A}^i\right)$ are chosen to be (unweighted) squared Euclidian distances respectively between V^i and $\hat{V}^i$, and between A^i and $\hat{A}^i$. This choice is largely discussed in Section IV. X (resp. $\hat{X}$) denotes the matrices that results from the double concatenation of V^i and A^i (resp. $\hat{V}^i$ and $\hat{A}^i$), both over line (video and audio components) and over columns according to index i. In our application, i is a time index of M consecutive sets of data. M greater than 1 leads to the general matrix quantization case and $M=1$ reduces to the vector quantization case. Finally, α is a factor that gives more or less relative importance to the visual or audio distortion measure, allowing to tune the quantizer to allocate more coding precision on the video vector or on the audio vector.

Now that the distances are defined, codebooks can be designed by using the Generalized Lloyd Algorithm (GLA) [26], which basically consists in alternating between two steps, starting with an arbitrary codebook of size L and a large training set of data:

Step 1) Assign each matrix X (time-sequences of vectors) of the training database to the nearest matrix codeword $\hat{X}_j$ in the codebook, that is the codeword that match the so-called "nearest neighbor condition": $D_{AV}\left(X,\hat{X}_j\right)\le D_{AV}\left(X,\hat{X}_k\right)$ for $k=1$ to L. The database is now partitioned into L (or fewer) cells.

Step 2) Calculate the centroid $\hat{X}_j$ of each cell j, that is the matrix that minimizes $E\left[D_{AV}\left(X,\hat{X}_j\right)\right]$ over all matrices X of cell j. The new set of centroids becomes the new codebook (this justifies the same notation $\hat{X}_j$ for step 1 and 2).

An iteration of step 1 and 2 can only improve the codebook, or at worst leave it unchanged. Thus, iterations are repeated until the mean distortion (averaged over the entire database) no more decreases, ensuring that a local minimum is attained. The "splitting algorithm" is used for initialization of the codebook [26], [27]: beginning with $L=1$, the unique codeword (which is the mean value of the database) is split in two by adding some small random value and the GLA iterations are applied. The splitting process (applied to each codeword) and the GLA iterations are repeated r times until $L=2^r$. The splitting process is also used to fix the problem of eventual empty cells that may occur during step 1 (the empty cell codeword is discarded and the maximum cardinal codeword is split).

It is important to note that, since $D_V\left(V^i,\hat{V}^i\right)$ and $D_A\left(A^i,\hat{A}^i\right)$ are squared Euclidian distances, $D_{AV}\left(X,\hat{X}\right)$ is also a squared Euclidian distance with the video components being weighted by α and the audio components being weighted by $(1-\alpha)$. In such case, the new codewords calculated during every "centroid calculation step" of the GLA algorithm are simply the means of the cells [27]. Moreover, since the mean vector/matrix is defined as a vector/matrix which each component is a mean value, each AV centroid of any AV cell is the concatenation of the V and A centroids. Therefore, for AV quantization, the audio-visual distance and the coefficient α are useless in the "centroid calculation step," which is identical to separate V and A centroids calculation. They are only used in the "nearest neighbor step" of the GLA, when the video and audio parameter sets are jointly quantized (i.e. classified) using D_{AV} rather than separately quantized using D_V and D_A.

D. Video to Audio Classified Vector Quantization

As mentioned before, multi-stage VQ/MQ consists of cascading several VQ/MQ blocks, with each following block being in charge of quantizing the quantization error at the output of the preceding block. Another form of cascading successive blocks is to be mentioned here, classified VQ/MQ. The principle is the following: the output of either a VQ/MQ or decision (or hybrid VQ/MQ-decision) block is used to drive the choice of the next block between several possible quantifiers. For example, this structure has been successfully used for differentiating the quantization of voiced and unvoiced speech LSPs [28], [29]. The main difference with multi-stage VQ/MQ is that several blocks are available at each stage, leading to a "tree structured" VQ/MQ, the blocks of different stage being possibly different in nature.

All these considerations make this structure particularly suitable for audio-visual coding since the audio-visual coherence should allow to "predict" the value of either visual or audio data from the other modality. Therefore, the quantized value of a set of parameters of one modality should allow to reduce the set of quantized possible values for the set of data of the other modality, thus allowing to select the "appropriate" quantizer. Therefore, we propose a classified VQ/MQ for jointly coding the audio-visual speech parameters, where the output of the first video VQ/MQ block drives the choice between several audio blocks (Fig. 3). Let us call this structure video-to-audio classified VQ/MQ (V-A C-VQ/MQ). For each audio-visual set of data, the video vector is first VQ/MQ quantized. Then the result of this r bits quantization determines which of the 2^r audio VQ/MQ is to be used to quantify the corresponding audio vector. A dual audio-to-video structure may also be considered but was not presented in this study for the purpose of presentation simplicity. Note that no audio-visual distance need to be defined for such structures, since both previously defined audio and visual distance, D_A and D_V, are separately used at each stage of the quantifier.

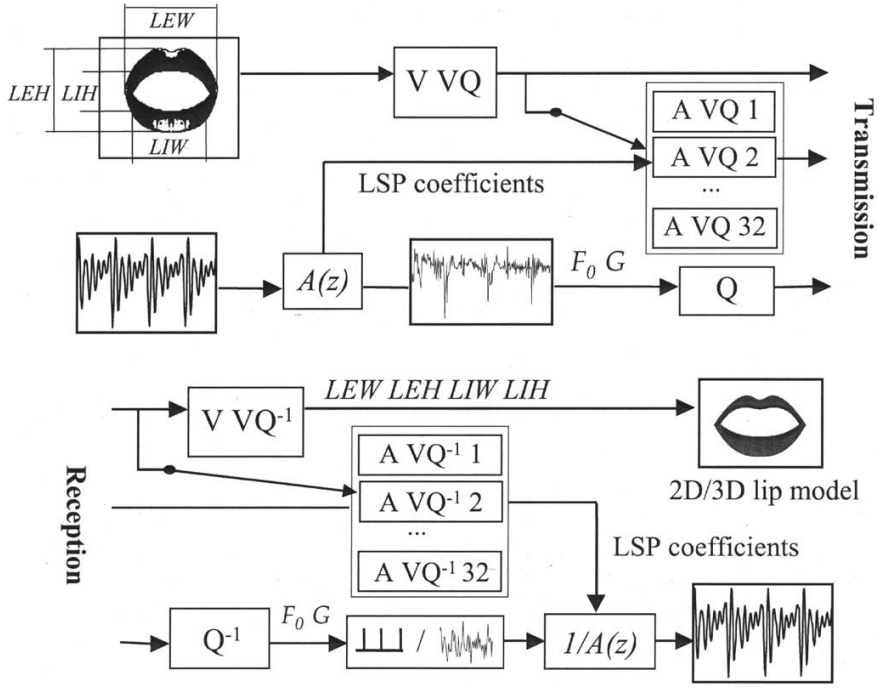

Fig. 3. Structure of an audiovisual speech codec based on classified video to audio vector quantization of lip and LSP parameters. Residual acoustic signal parameters (e.g., gain G and fundamental frequency F_0) are quantized and transmitted separately.

E. Important Remark

Generally, the elaboration of sub-optimal (compared to single-stage VQ/MQ) multi-stage or classified VQ/MQ is mainly justified by the need to reduce the complexity codebook generation and search procedures. This suggests an important remark concerning the design of VQ/MQ for the audio-visual data and more particularly, single-stage audio-visual VQ/MQ. In the literature, high quality LSP quantization is generally reported to be obtained for single stage VQ at around 20 bits/frame (or multi-stage or split VQ at around 25 bits/frame.) In the experiment section, we present results of audio-visual single-stage VQ/MQ obtained from 5 to 14 bits/frame, thus leading to relatively poor quality. Two remarks can be made to give a realistic aspect to this study regarding this point. First, audio coders can be elaborated with relatively low resolution LSP vector/matrix quantizers (around 10 bits/frame), since they are able to provide limited quality but still quite intelligible speech [29]. MQ has been applied more particularly to such objective [24]. Second, the audio-visual single stage VQ/MQ that is proposed in this paper can be regarded and used as a first stage for a multi-stage VQ/MQ structure, the following either audio-only, video-only or audio-visual stages (not considered here) providing complementary quality. With such structures, audio-visual VQ/MQ can always benefit to audio-visual speech compression if the first block can provide better results than any audio VQ/MQ plus video VQ/MQ separate blocks. Note that the same remark can be made for the blocks of the proposed V-A classified VQ/MQ: each cascaded video or audio blocks can also be easily ("separately") extended to multi-stage VQ/MQ if the bit rates proposed in the experimental sections cannot ensure desired quality.

III. Computer Simulation Results

A. Corpus and Training Procedures

VQ/MQ were designed using a nearly 3 h duration training database of audio-visual speech signals produced by eight French native speakers (four male and four female). The corpus consisted of continuous speech read from randomly chosen French press articles "as naturally as possible." The audio signal were sampled at 8 kHz, and a 10th order LPC analysis (autocorrelation method) was carried out every 20 ms with a 20 ms duration (nonoverlapping) sliding window, so that the center of the window corresponded to the lip parameters extraction time instant (remember that the video period is 20 ms). Thus, every 10 LSP parameters set was paired with the corresponding $[LIW\ LIH\ LEW\ LEH]$ set. The total amount of such audio-visual sets was about 510 000. For the MQ, these audio-visual vectors were gathered in sets of M consecutive vectors ($M = 2$ or 3) with a "sliding" gathering so that all the parameter trajectories present in the corpus should be used in the training process and the approximately identical (maximum) amount of data could be used for VQ and MQ (if N vector sets are available for VQ training, then $N-M+1$ gathered vector sets are available for MQ training with the sliding technique, with $N \approx 510\,000$ and $M = 2$ or 3).

All VQ/MQ codebooks were designed using the GLA algorithm with the "splitting" technique for initialization as described in Section II-C. The audio-visual distance D_{AV} of (1) was used to design AV-SS-VQ/MQ, while the D_A and D_V distances were separately used to design the A-SS-VQ/MQ and V-SS-VQ/MQ that were used as reference for performance

TABLE I
COMPARED RESULTS OF VIDEO, AUDIO AND AUDIO-VISUAL SINGLE STAGE VECTOR/MATRIX QUANTIZATION. M DENOTES THE NUMBER OF COLUMNS FOR THE MATRIX QUANTIZATION. V DENOTES THE MEAN VIDEO DISTORTION MEASURES (IN MM), AND A DENOTES THE MEAN AUDIO DISTORTION MEASURES (IN $10^{-2} \cdot$ rad)

	resolution (bits)		5	6	7	8	9	10	11	12	13	14
	Separate	V	1.565	1.245	1.004	0.819						
		A	7.755	7.044	6.451	5.925	5.459	5.031	4.643	4.272	3.909	
	AV	V			2.294	2.045	1.843	1.700	1.557	1.431	1.320	1.212
M=1	α= 0.1	A			7.768	7.190	6.666	6.242	5.804	5.389	4.980	4.562
(VQ)	AV	V			1.876	1.663	1.489	1.343				
	α= 0.2	A			8.491	7.831	7.232	6.705				
	Separate	V	1.884	1.560	1.305	1.099						
		A	8.713	8.102	7.566	7.154	6.737	6.357	6.003	5.668	5.334	
	AV	V			2.509	2.270	2.077	1.905	1.757	1.625	1.501	1.385
M=2	α= 0.1	A			8.873	8.345	7.839	7.387	6.963	6.558	6.162	5.759
(MQ)	AV	V			2.082	1.864	1.698					
	α= 0.2	A			9.487	8.890	8.314					
	Separate	V	2.124	1.768	1.494	1.272	1.090					
		A	9.370	8.797	8.283	7.881	7.483	7.101	6.740	6.399	6.060	
M=3	AV	V			2.661	2.416	2.219	2.077	1.920	1.773	1.637	1.508
(MQ)	α= 0.1	A			9.533	9.003	8.491	8.102	7.658	7.244	6.834	6.417
	AV	V			2.250	2.036	1.844					
	α= 0.2	A			10.213	9.609	9.068					

evaluation of the AV structures, as well as the A and V blocks of the classified VQ/MQ. It is interesting to note that in the presented results, all the 510 000 vectors of the corpus were used both for training of the quantizers and testing (this means that the distortions that are presented are the ones that were obtained at the end of the GLA algorithms). This should be generally avoided but it was verified here that the results obtained were very close to the results obtained on 50 000 additional testing vectors (about 6,000 per speaker) that were not used for training ("very close" means that the relative differences between the distortion values obtained in the two cases were always less than 2% and that all the comparative results discussed in the next section were identical). This shows that the training procedure on such a large amount of data ensures sufficient generality to the results.

The results of the different quantization processes are all given in terms of video and audio root mean square distortion measures, that are the roots of D_V and D_A averaged over the entire training database (averaging was made on squared distances before taking the root of the result).[4] The audio distortions are given in rad, which is a usual dimension for the LSP parameters (when normalized between 0 and π). The video measures and distortions are expressed in mm both to provide a realistic articulatory-related quantitative evaluation and to ensure homogenous range of values between the different speakers. For this aim, a ruler was captured at the beginning of each video recording at a controlled distance of the camera and provided a reference for pixel to mm conversion (the head of the speaker could not move during the recordings). Note that all distortion measures were normalized by the dimension of the vectors or matrices considered to ensure direct comparison of the results for each video or audio modality independently from quantization condition.

[4]In the AV-SS quantization process of each AV vector, the D_V and D_A terms of D_{AV} were "isolated" for averaging.

B. AV-SS-VQ/MQ Results

We first present in this section the tuning of the α parameter. Then the removing of inter-frame correlation with MQ compared to VQ is briefly discussed. This is not an original result of the study but it can be mentioned as audiovisual VQ and MQ are both implemented. Finally, comparison between AV and separate A and V quantization is largely described.

1) Tuning of α: The results of the AV-SS-VQ/MQs are presented in Table I. They cover two values of α and bit rate ranging from 5 to 14 bits. The value $\alpha = 0.1$ was chosen after pilot experimentation aiming to "equilibrate" in some sense the V and A distances of the AV VQ/MQ in reference with the V and A distances obtained with separate V and A VQ/MQ: the aim was to select a value that would allow comparable results between the AV quantization from around 10 to 14 bits and arbitrary V and A quantization over 5 bits. For example, a 5 bits V-SS-VQ/MQ can be considered as a minimum size codebook reference because it corresponds to a set of 32 lip shapes while it was shown in [30] that French could be characterized visually by a set of 23 different lip shapes. Besides, 5 to 10 bits is a standard range for individual blocks of multi-stage LSP VQ/MQ. This equilibrium between A and V coding accuracy is arbitrary and finally user dependent. Results are also given for $\alpha = 0.2$ to illustrate the influence of this factor but are within a limited bit range because of the high time consumption of the experiments (e.g., 3 days of computing are necessary for a 11 bits VQ on the machine that was used, an intel PIII bi-processor clocked at 850 MHz).

2) Removing of Inter-Frame Correlation With MQ: The results obtained with the MQ confirm that this method efficiently exploits the speech inter-frame correlation, as previously observed in [24], [25]. For example, let us first consider separate V and A MQ. We can see in Table I that a 2-columns V MQ with respectively 7 and 8 bits lead to significantly lower averaged video distortion than a V VQ with, respectively, 5 and 6 bits (even 6 bit is sufficient to the 2 columns MQ to equal the 5 bits VQ performance). This means that only 2 supplementary

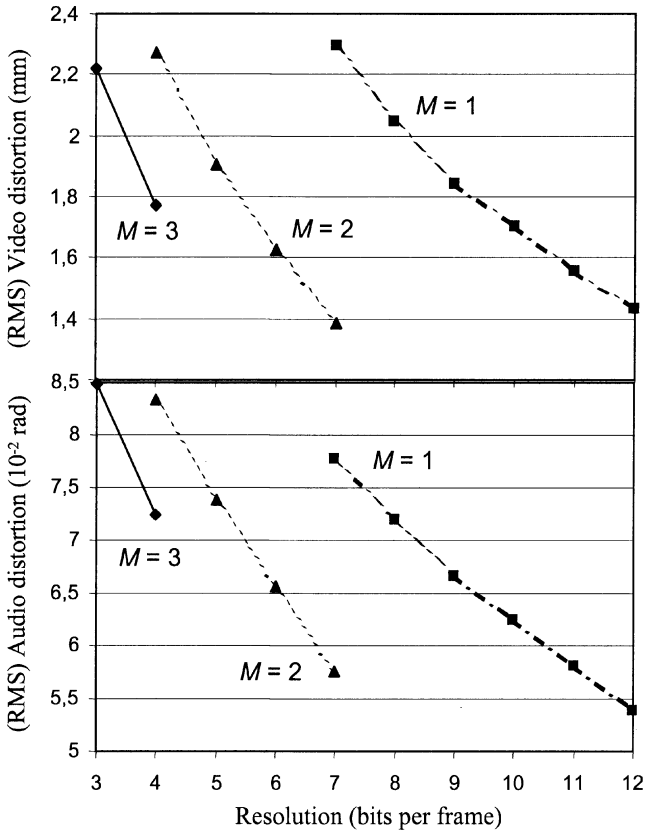

Fig. 4. Video (top) and audio (bottom) root mean square distortion measures for audio-visual (AV) ($\alpha = 0.1$) single stage vector ($M = 1$) and matrix ($M = 2$ and 3) quantization as a function of resolution. For easier comparison, the resolution is defined in this figure as the mean number of bits used to encode one set of AV parameters (before reporting selected distortion values of Table I in the figure, the corresponding resolution values of Table I were divided by M).

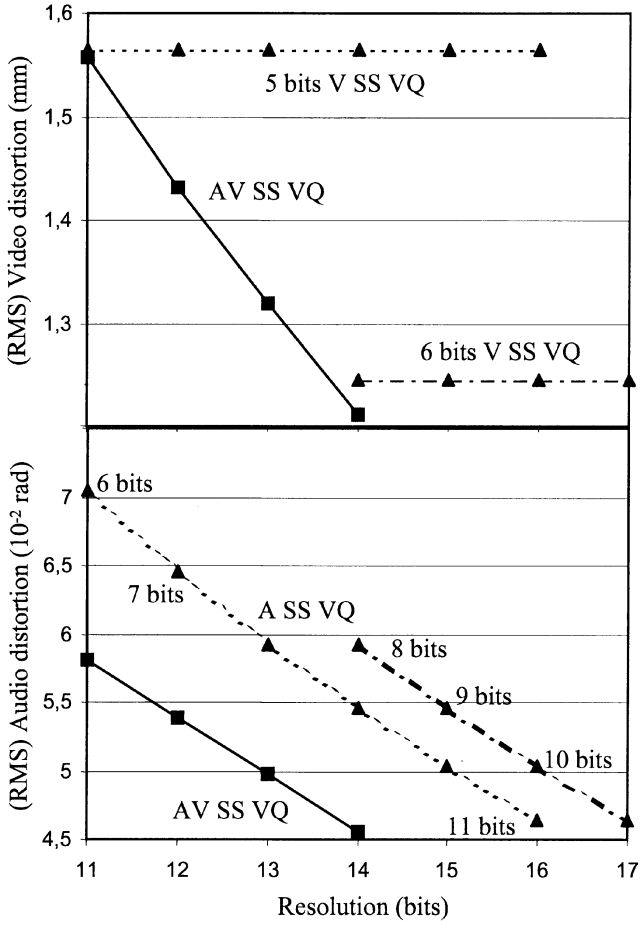

Fig. 5. Video (top) and audio (bottom) root mean square distortion measures for audio-visual (AV) ($\alpha = 0.1$) versus separate video (V) + audio (A) single stage vector quantization as a function of the total resolution (A resolution + V resolution for the separate quantizers; the separate A and V resolutions are given in the figure). The values are extracted from Table I. The dashed (resp. dotted) curb of the V separate VQ is associated with the dashed (resp. dotted) curb of the A separate VQ to provide a total resolution identical to the AV VQ.

bits are necessary to encode a supplementary V vector when the two vectors are coded together with the MQ, allowing to save 1.5 or 2 bits per frame. In the same manner, the 2-columns A MQs at respectively 10, 12 and 13 bits lead to better performances than A VQs with respectively 7, 8, and 9 bits (allowing to save respectively 2, 2, and 2.5 bits per vector). These results can be generalized to the AV case (see Fig. 4). For example the series of 2-columns AV MQs ($\alpha = 0.1$) at respectively 11, 12, 13, and 14 bits lead to better V and A performance than the corresponding AV VQs at respectively 8, 9, 10, and 11 bits (allowing to save respectively 2.5, 3, 3.5, and 4 bits per frame). The results obtained with the 3-column MQs show that inter-frame correlation can be efficiently exploited within 3 consecutive vectors: for example, the 3-columns AV MQ at 14 bits ($\alpha = 0.1$) gives lower A distortion and quite lower V distortion than the corresponding AV VQ ($\alpha = 0.1$) at 9 bits, allowing to save respectively 13 bits every 3 vectors (that is 4.33 bits per frame) while providing significant video coding gain. This is the best result obtained by MQ compared to 9 bits VQ.[5]

Note that, compared to VQ with the same value of α, MQ seems to have slightly changed the balance between V and A distortion to the benefit of V and this unexpected effect is more pronounced when increasing the number of columns. This suggests that V data may be more likely to benefit from inter-frame correlation than A data in a general manner (see the separate V and A results at the beginning of this section) and in an AV joint coding process, which may be useful information to be considered for further development of this application.

3) AV versus V and A Separate Quantization: Let us now deal with the very heart of the study, that is the comparison between AV and separate V and A VQ/MQ. The results of Table I show the efficiency of the AV processing. For example, for $\alpha = 0.1$, an AV VQ at 11 bits provides both V and A distortions that are lower than the ones obtained respectively with a 5 bits V VQ and an 8 bits A VQ, thus allowing to save 2 bits out of 13 (15.4% gain) while providing more accurate coding (see also Fig. 5). Exactly the same result is obtained with the 2-columns MQs (Fig. 6). The bit saving can be increased if we increase the resolution of each quantizer. For example, the 14 bits AV VQ gives lower V and A distortions than the 6 bits V VQ and the 11 bits A VQ, leading to 3 bits saving out of 17 (17.7% gain) while providing more accurate coding. With the 2-columns MQ, the

[5]In what concerns audio data, [25] reported an optimal value of 4 for M, while [24] concluded that with this value, "vowel-like sound were well captured while consonant were less well captured." In the present study, M was not greater than 3.

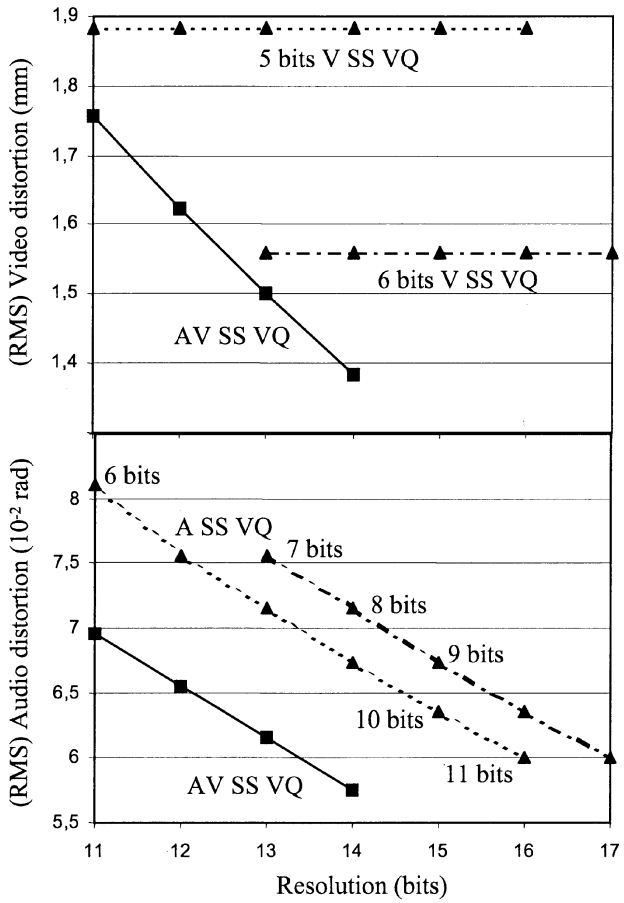

Fig. 6. Video (top) and audio (bottom) root mean square distortion measures for audio-visual (AV) ($\alpha = 0.1$) versus separate video (V) + audio (A) single stage 2-columns matrix quantization as a function of the total resolution (A resolution + V resolution for the separate quantizers; the separate A and V resolutions are given in the Figure). The values are extracted from Table I. The dashed (resp. dotted) curb of the V separate MQ is associated with the dashed (resp. dotted) curb of the A separate MQ to provide a total resolution identical to the AV MQ.

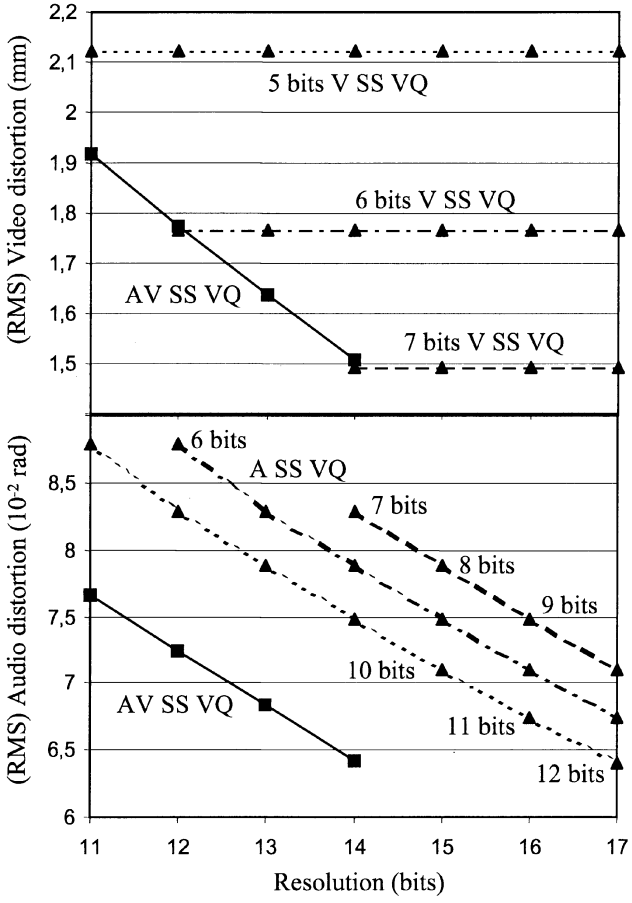

Fig. 7. Video (top) and audio (bottom) root mean square distortion measures for audio-visual (AV) ($\alpha = 0.1$) versus separate video (V) + audio (A) single stage 3-columns matrix quantization as a function of the total resolution (A resolution + V resolution for the separate quantizers; the separate A and V resolutions are given in the figure). The values are extracted from Table I. The dashed (resp. dotted and dashed/dotted) curb of the V separate VQ is associated with the dashed (resp. dotted and dashed/dotted) curb of the A separate VQ to provide a total resolution identical to the AV VQ.

coding gain is extended to 3 bits out of 16 (18.8%): the AV MQ with 13 bits is more efficient than separate V and A MQs at respectively 6 and 10 bits. When the number of bits of the AV MQ is increased from 11 to 12 or from 13 to 14, no supplementary bit can be saved: the results can be advantageously compared with V and A distortions from separate MQs with respectively 5 and 9 bits in the first case, and with respectively 6 and 11 bits in the second case. This can be explained by the fact that the decreasing of the A distortion in the AV coding is regular in the sense that this A distortion is always lower in the AV coding with r bits than in the A coding with $r - 3$ bits, while the V distortion is not as much regular. But even if no supplementary bit can be saved from 11 to 12 bits or from 13 to 14 bits, the fact that the gain for V distortion increases always demonstrates the advantage of the AV joint coding over separate V and A coding.

For $\alpha = 0.2$, more coding precision is given to the V coefficients, leading to a lower V distortion and a higher A distortion than in the $\alpha = 0.1$ condition for each tested resolution. At the same time, the bit saving appears quite similar (though not all resolutions could be tested): for example, for $\alpha = 0.2$, an AV VQ with 9 bits provides both V and A distortions that are lower than the ones obtained respectively with a 5 bits V VQ and a 5 bits A VQ. Thus, the bit saving is here 1 out of 10, but the distortion gains are quite notable, especially for the A distortion (about 7% from 0.077 55 to 0.072 32 rad). Again, the same result is obtained with the 2-columns MQ, while the gains on V and A distortion are in this case more "equilibrated."

Finally, the advantage of jointly coding V and A data is confirmed in the 3-columns MQ case (see Fig. 7), which is the more efficient condition according to the inter-frame correlation removing criterion. For example, the 3 columns AV MQ at 14 bits ($\alpha = 0.1$) provides quite similar (while no lower) V and A distortions than, respectively, the 3 columns V MQ at 7 bits (1.508 mm versus 1.494 mm) and the 3 columns A MQ at 12 bits (0.064 17 rad versus 0.063 99 rad), leading to similar coding accuracy while saving 5 bits out of 19 (26.3% gain). This represents the best result in term of bit saving, but the condition on the distortions was slightly relaxed (no strict lowering) compared to the VQ case where 3 bits out of 17 were saved with V and A distortions significantly decreased. Note that if we strictly apply the distortion comparison criterion, the "3 bits out of 17" result of VQ can be generalized to 2 or 3 columns MQs ($\alpha = 0.1$)

since the 2 or 3 columns AV MQs at 14 bits both provide lower V and A distortion than corresponding separate V MQ at 6 bits and A MQ at 11 bits (see Figs. 6 and 7).

4) V to A Classified VQ/MQ: The video-to-audio classified VQ/MQ of Section II-D was implemented. The 5 bits V-SS-VQ of the preceding section was used as the first block. The complete set of about 510 000 audio-visual vectors were quantized with this block (that is only the video components were quantized) leading to a cluster of $2^5 = 32$ classes of audio-visual data, each class being associated to a video centroids. Then, the audio components of each class were used to generate an A-SS-VQ by using the GLA algorithm with splitting initialization. Finally, A-SS-VQs were obtained, each of them being associated with one of the 32 video centroids for the cascading quantization process. This whole design was repeated for MQ, that is V and A MQ blocks were also cascaded, with $M = 2$ and $M = 3$.

One major problem that is usually encountered during the design of such structure is the limitation due to the amount of data needed. Indeed, the 510 000 audio sets of the data where unequally affected within the 32 classes, leading to subsets of around 3000 (for the worst case) to around 33 000 (for the max) sets of data. Thus the audio resolution had to be limited in each class to ensure reliable centroïd generation. A criterion for such validity is to ensure that the ratio R between the number of data in a training set and the number of centroïds to be generated is greater than a given threshold $Rmin$, typically greater than one hundred. Once $Rmin$ is fixed, it determines the resolution of the audio VQ/MQ in each class as the maximum number of bits so that $R > Rmin$ for that class. Thus, the resolution can be different from one class to another (from 4 to 7 bits with the values of $Rmin$ considered), leading to a variable rate quantizer. In such case, r denotes the mean resolution, that is averaged over the different classes by using the cardinals of the classes as weights.

The results obtained with the V-A C-VQ/MQ are presented in Table II, still in terms of root mean square audio distances (the video distortion measures are in Table I). Note that the values of $Rmin$ were carefully chosen so that either the resolution or the audio distortion had a value as close as possible as one of the values of the A VQ/MQ corresponding condition of Table I to quantify easily the improvement due to the cascaded V-A structure.

The results of Table II show that again, either better audio distortion or resolution (or even both) can be obtained with the cascaded V-A C-VQ/MQ structure compared to the V and A separated VQs/MQs. Indeed, for any given identical resolution, the audio distortion obtained with the cascaded structure is significantly lower than the one obtained with the separated audio VQ/MQ. This is verified for each value of M. For example, for $M = 1$ and for $r = 5$, 6 and 7 bits (rigorously 5.01, 6.01 and 7.01, respectively), the audio distortions are respectively 10.6%, (0.069 33 rad versus 0.077 55 rad), 11.2% (0.062 58 rad versus 0.070 44 rad) and 9.7% (0.058 27 rad versus 0.064 51 rad) lower. These results are slightly less impressive for $M = 2$: respectively 8.5%, 9.2% and 8.0% gains on audio distortions are obtained at respectively 5, 6 and 7 bits. For $M = 3$, we obtain respectively 7.5%, 8.6% and 7.8% which represent a similar range

TABLE II
RESULTS OF CLASSIFIED VIDEO TO AUDIO VECTOR/MATRIX QUANTIZATION. M DENOTES THE NUMBER OF COLUMNS FOR THE MATRIX QUANTIZATION, r IS THE RESOLUTION AND $Rmin$ THE RATIO OF THE NUMBER OF DATA AND CENTROIDS (SEE THE TEXT FOR DETAILS). THE MEAN AUDIO DISTORTION MEASURES ARE GIVEN IN $10^{-2} \cdot \mathrm{rad}$

M=1	r (bits)	5.01	5.65	6.01	6.56	7.01
	$Rmin$	370	225	185	125	90
	$\sqrt{\overline{D_A}}$	6.933	6.440	6.258	5.920	5.827
M=2	r	5.01	5.68	6.00	6.39	7.00
	$Rmin$	405	255	195	150	98
	$\sqrt{\overline{D_A}}$	7.975	7.545	7.358	7.150	6.959
M=3	r	5.01	5,63	6.01	6.30	7.01
	$Rmin$	400	256	195	160	98
	$\sqrt{\overline{D_A}}$	8.667	8.262	8.039	7.880	7.635

of values. Thus, for the V-A classified structure, the VQ seems to provide slightly better "classified versus separate" gains than the MQ, which provides quite similar gains for $M = 2$ and $M = 3$, although the 3 columns MQ was the more efficient structure in the single stage structure of the previous sections.

Besides, it is interesting to note that the audio distortions obtained with the classified VQ/MQ are well greater than the ones obtained with the (optimal) audiovisual single stage VQ/MQ with the same overall bit rate. For example, the above mentioned value of 0.058 27 rad for the V-A C-VQ at 7 bits is to be unfavorably compared with 0.053 89 rad which is obtained with the AV-SS-VQ at $5+7 = 12$ bits (9.7% difference). Other similar difference can be reported for other bit rates and values of M, confirming that the classified structure provides intermediate performance between the separate V and A structure and the optimal single stage AV structure, from the distortion criterion.

The non entire values of r allow to quantify the gain of the cascading structure in terms of bit rate. For example, for $M = 1$ (VQ), similar (while always lower) audio distortions are obtained with the V-A classified structure at respectively 5.65 and 6.56 bits and with the A separated VQ at respectively 7 and 8 bits, leading to respectively 19.3% and 18% audio bit savings. A gain of 20.1% is obtained for $M = 2$, from 8 bits in separate A MQ to 6.39 bits in V-A C-MQ (we have also 18.9% gain from 7 to 5.68 bits). Finally, the best gain (21.3%) is obtained for $M = 3$, from 8 bits in separate A MQ to 6.30 bits in V-A C-MQ (we have also 19.6% gain from 7 to 5.63 bits). Surprisingly, the gains in resolution are slightly increasing with M, while the V-A C-VQ provided the best gains of audio distortion measure. However, this may be due to the variability of the rate-distortion relation when estimated from data, together with the fact that the presented results are relative gains calculated from two estimated measures, and given also that the different values of M provide quite similar results.

Note that these gains in resolution should be tempered by including the 5 video bits of the first block of the classified quantizer, which is identical in both classified and separate structures. If this is done, the resolution gains are close to 11% depending on the resolution and M, showing again the intermediate status of the classified structure between the separate and

single-stage structures regarding the distortion-rate criterion. In a general manner, this latter point highlights the global coherence of the different results and leads to expect the AV coherence to largely benefit to most coder structures based on vector or matrix quantization.

IV. Concluding Remarks

We have presented a joint audio and visual quantizer for low bit rate audiovisual speech coding. This quantizer was based on single-stage vector/matrix quantization and the definition of an audiovisual distance. It allowed to simultaneously quantize face parameters which were, in this study, lip shape parameters and LPC-type audio parameters which were usual LSP parameters. Compared to separate audio and video quantizers build in the same configuration (except audio and video data separation), the audiovisual quantizer allows significant gains in terms of video and audio mean distortion measures and/or resolution. In term of bit saving, the more general result appears to be "3 bits saved out of 17" which is obtained for the three tested values of M. In each case, both video and audio mean distortion measures are assumed to be lower with the AV joint quantization than with the separate V and A quantizations. If this condition is slightly relaxed so that video and audio distortion are "quite identical" in the AV and separate V and A conditions, the bit saving can reach 5 out of 19. This result was obtained for a 3-column AV MQ, which was shown besides to efficiently exploit inter-frame correlation and thus can be considered as the most efficient quantizer we obtained.

In order to reduce the complexity of the overall process, a classified tree-like structure was also proposed, consisting of two cascaded video and audio quantization blocks. It has been shown to provide "intermediate" distortion-rate performances between the optimal single-stage AV structure and separated V and A structures.

Although these results are quite promising, further experiments need to be conducted to consolidate our first approach. Among them, the problem of the distortion measures comes in first position, in relation with perceptual evaluation of the process. In this study, quite simple (unweighted Euclidian) audio and video distances were used and linearly combined. The main reason for this choice was that, as mentioned before, the main objective of this study was to conduct a first quantitative comparison of joint audio-visual quantization versus audio and visual separate quantization under the same configuration, and not to tune more efficient (thus sophisticated) configurations for the different elements of the quantizers. However, now that this goal is achieved, we can think of using more sophisticated and performing distortion measures involving perceptual considerations. Concerning the audio data, different kinds of weighted distances (e.g., in [18], [19], [25]) were shown to provide more satisfying relationship between quantization accuracy of LSP parameters, spectral distortion measures which are usually considered in audio speech coding (e.g., root mean square difference between log-power LPC spectra) and perceptual judgment. Such distances may be considered in the future extensions of this work. Similarly, the video distance D_V may also be improved by taking into account the perceptual weight of each video parameter, their possible range of value and adapted scale (e.g., give more importance to the quantization of small values of the lip parameters), and including the fact that the lip parameters should be extended to other face descriptors. Note that the video distance that was used in this preliminary study is nevertheless intrinsically related to perceptual considerations since it is applied on perceptually crucial parameters and the Euclidian distance directly involves a geometric interpretation.

In direct relation with the distortion measure problem, the subjective evaluation of the joint coding process must be considered in the near future. This involves the elaboration of a complete multi-speaker analysis-modeling-coding-synthesis system and its complete assessment including intelligibility tests (e.g., audiovisual identification of noisy stimuli) and subjective quality tests in addition to distortion-rate measures. This represents a huge task since it is currently motivating a series of works on (among others) efficient automatic face analysis, head model speaker normalization/conversion, and relation of the proposed process with MPEG-4 coding techniques. Besides, the protocols for subjective evaluation of the system may take a particular aspect in the specific joint audiovisual context of this study: subjective evaluation should include the study of the perceptual consequences of the joint coding process, e. g. the "equilibrium" given by the weight α between audio and video parameter trajectories accuracy. It is interesting to note for example that [25] reported a superiority of MQ over VQ for audio subjective measure due to the "smooth trajectories" involved by the averaging process during the MQ codebook design. This property that has been found to be perceptually important when synthesizing high quality audio speech should be taken into account in the audiovisual case where visual trajectories are also involved and could be smoothed as well.

To finish with an encouraging remark, it can be mentioned here that very preliminary informal tests were conducted with synthesis lip models animated with parameters that were quantized by different VQs/MQs presented in this study. They cannot be presented in details in this paper but they provided satisfying preliminary results, with good general lip movements/sounds synchronization and no gross artifacts. This is especially encouraging given the relatively poor resolution range used in this study (remember that the quantizers can be regarded as the first blocks of multi-stage quantizers). Altogether, the encouraging results obtained in this work may, to our opinion, give a new impulse to the relationship between speech coding and audiovisual speech synthesis.

References

[1] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lipreading*, B. Dodd and R. Campbell, Eds. London, U.K.: Lawrence Erlbaum, 1987, pp. 3–51.

[2] L. E. Bernstein and C. Benoît, "For speech perception by humans or machines, three senses are better than one," in *Proc. ICSLP*, 1996, pp. 1477–1480.

[3] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, pp. 23–43, 1998.

[4] D. G. Stork and M. Hennecke, Eds., *Speechreading by Man and Machine: Models, Systems and Applications*. Berlin, Germany: Springer-Verlag, 1996.

[5] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. Amer.*, vol. 109, pp. 3007–3020, June 2001.
[6] P. Doenges, T. K. Capin, F. Lavagetto, J. Ostermann, I. Pandzic, and E. Petajan, "MPEG-4: Audio/video and synthetic graphics/audio for real-time, interactive media delivery," *Image Commun. J.*, vol. 9, no. 4, pp. 433–463, 1997.
[7] A. M. Tekalp and J. Ostermann, "Face and 2D mesh animation in MPEG-4," *Signal Process.: Image Commun.*, vol. 15, pp. 387–421, 2000.
[8] G. Bailly, "Audiovisual speech synthesis," in *Proc. Eur. Tutorial and Research Workshop on Speech Synthesis*, Perthshire, U.K., 2001.
[9] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *Proc. Audio-Visual Speech Processing Workshop*, Aalborg, Denmark, 2001, pp. 90–97.
[10] R. Rao and T. Chen, "Cross-modal predictive coding for talking head sequences," in *Proc. ICASSP*, 1996, pp. 2058–2061.
[11] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Mag.*, vol. 18, pp. 9–21, Jan. 2001.
[12] M. Lallouache, "Un poste 'visage-parole'. Acquisition et traitement de contours labiaux," in *Proc. XVIII J. d'Études sur la Parole*, Montréal, QC, Canada, 1990, pp. 282–286.
[13] L. Revéret and C. Benoît, "A new 3D lip model for analysis and synthesis of lip motion in speech production," in *Proc. Audio-Visual Speech Processing Workshop*, Sydney, Australia, 1998, pp. 207–212.
[14] B. L. B. Goff, T. Guiard-Marigny, and C. Benoît, "Read my lips… and my jaw! How intelligible are the components of a speaker's face?," in *Proc. Eur. Conf. Speech Commun. Tech.*, Madrid, Spain, 1995, pp. 291–294.
[15] B. Le Goff, T. Guiard-Marigny, and C. Benoît, "Analysis-synthesis and intelligibility of a talking face," in *Progress in Speech Synthesis*, J. P. H. Van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds. New York: Springer-Verlag, 1996, pp. 235–244.
[16] F. Elisei, G. Bailly, M. Odisio, and P. Badin, "Clones parlants 3D video-réalistes: Application à l'interprétation de FAP MPEG-4," in *Proc. Compression et Représentation des Signaux Audiovisuels*, Dijon, France, 2001, pp. 145–148.
[17] N. Sugamura and F. Itakura, "Speech analysis and synthesis method developed at ACL in NTT-From LPC to LSP," *Speech Commun.*, vol. 5, pp. 199–215, 1986.
[18] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 3–14, Jan. 1993.
[19] J. Pan and T. R. Fischer, "Vector quantization of speech line spectrum pair parameters and reflection coefficients," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 106–115, 1998.
[20] P. Hedelin, "Single stage spectral quantization at 20 bits ," in *Proc. IEEE ICASSP*, 1994, pp. 525–528.
[21] N. Phamdo, N. Favardin, and T. Moriya, "Combined source-channel coding of LSP parameters using multi-stage vector quantization," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, 1991, pp. 36–38.
[22] W. P. LeBlanc, B. Battacharya, S. Mahmoud, and V. Cupperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 373–385, Oct. 1993.
[23] M. Yong, G. Davidson, and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction," in *Proc. IEEE ICASSP*, 1988, pp. 402–405.
[24] C. Tsao and R. M. Gray, "Matrix quantizer design for LPC speech using the generalized Lloyd algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 3, pp. 537–545, 1985.
[25] C. S. Xydeas and C. Papanastasiou, "Split matrix quantization of LPC parameters," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 2, pp. 113–125, 1999.
[26] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–94, 1980.
[27] R. M. Gray and A. Gersho, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers, 1992.
[28] R. Hagen, E. Paksoy, and A. Gersho, "Voicing-specific LPC quantization for variable-rate speech coding," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 5, pp. 485–494, 1999.
[29] D. Y. Wong, B. H. Juang, and A. H. Gray, "An 800 bit/s vector quantization LPC vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 5, pp. 770–780, 1982.
[30] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry, "A set of visual french visemes for visual speech synthesis," in *Talking Machines: Theories, Models and Designs*, G. Bailly, C. Benoît, and T. R. Sawallis, Eds. Amsterdam, The Netherlands: Elsevier, 1992, pp. 485–504.

Laurent Girin received the M.Sc. and the Ph.D degrees in signal processing from the Institut National Polytechnique de Grenoble, France, in 1994 and 1997, respectively.

In 1997, he joined the Ecole Nationale d'Electronique et de Radioélectricité de Grenoble, where he is currently an Associate Teacher in electrical engineering and signal processing. His research activity takes place in the Institut de la Communication Parlée (Speech Communication Lab) in Grenoble. His current research interests are in audiovisual speech processing with application to speech coding, speech enhancement and audio/speech source separation.